Psychometrika

A JOURNAL DEVOTED TO THE DEVELOPMENT OF PSYCHOLOGY AS A QUANTITATIVE RATIONAL SCIENCE

THE PSYCHOMETRIC SOCIETY · ORGANIZED IN 1935

VOLUME 9
NUMBER 4
DECEMBER
1 9 4 4

PSYCHOMETRIKA, the official journal of the Psychometric Society, is devoted to the development of psychology as a quantitative rational science. Issued four times a year, on March 15, June 15, September 15, and December 15.

DECEMBER 1944, VOLUME 9, NUMBER 4

Offices of Publication: 23 West Colorado Avenue, Colorado Springs, Colorado. Entered as second class matter, September 17, 1940 at the Post Office of Colorado Springs, Colorado, under the act of March 3, 1879. Editorial Office, 2539 Briarcliffe Avenue, Cincinnati, Ohio.

Subscription Price:

To non-members, the subscription price is $5.00 per volume of four issues. Members of the Psychometric Society pay annual dues of $5.00, of which $4.50 is in payment of a subscription to *Psychometrika.* Student members of the Psychometric Society pay annual dues of $3.00, of which $2.70 is in payment for the journal.

The subscription price to libraries and other institutions is $10.00 per year; this price includes one extra copy of each issue.

Applications for membership and student membership in the Psychometric Society should be sent to

THELMA GWINN THURSTONE
Chairman of the Membership Committee
University of Chicago
Chicago, Illinois

Membership dues and subscriptions are payable to

IRVING D. LORGE
Institute of Educational Research
Columbia University
New York, New York

Articles on the following subjects are published in *Psychometrika*:

(1) the development of quantitative rationale for the solution of psychological problems;

(2) general theoretical articles on quantitative methodology in the social and biological sciences;

(3) new mathematical and statistical techniques for the evaluation of psychological data;

(4) aids in the application of statistical techniques, such as monographs, tables, work-sheet layouts, forms, and apparatus;

(5) critiques or reviews of significant studies involving the use of quantitative techniques.

(*Continued on the back inside cover page*)

Psychometrika

CONTENTS

VOLUME NINE DECEMBER 1944 NUMBER FOUR

PSYCHOMETRIKA—VOL. 9, NO. 4
DECEMBER, 1944

STATISTICAL PROBLEMS IN THE EVALUATION OF ARMY TESTS

CYRIL BURT
DEPARTMENT OF PSYCHOLOGY, UNIVERSITY COLLEGE, LONDON

The introduction of psychological tests for personnel selection in the British forces has given rise to several novel problems in statistical procedure. The solutions proposed are in the main extensions of devices already familiar in educational psychology. The more important are: (i) where the criterion yields a threefold classification only, a method of triserial correlation or of biserial correlation assuming point-distributions for the extremes; (ii) where the data on which validation has to be based are drawn from a selected sample, a simplified form of Pearson's equations to correct for selection; (iii) where the best line of demarcation has to be deduced from theoretical rather than practical considerations, a formula based on the principle of minimal discrepancy.

In a paper recently prepared for the British Psychological Society (10), I have briefly outlined, so far as war-time conditions allow, the methods adopted for personnel selection in the British fighting services. Their elaboration has entailed much preparatory research, which in turn has given rise to new problems of more general interest. Broadly speaking, in the preliminary investigations on the collection and validation of appropriate tests,* the plan pursued has been much the same as that adopted for earlier studies in this country on guidance and selection within the educational and industrial fields. A typical inquiry involves five or six successive steps: (a) a preliminary job-analysis; (b) a tentative selection or construction of suitable tests; (c) a comparative study of the predictive value of each of the tests, in varying combinations, by means of partial regression (or some equivalent procedure); (d) a final standardization of a practicable battery, made up of the tests showing the highest predictive value, and supplemented by a specification of borderlines or norms for the various jobs; (e) a subsequent follow-up of the men tested and selected, in order to check, and (if necessary) refine still further, the proposed procedure; (f) where different jobs or training recommendations have to be allotted on the basis of one and the same set of

* In what follows, to save continual circumlocution, I shall use the word "test" to cover scores based on any empirical method of assessment, e.g., gradings derived from observations, interviews, questionnaires, rating-scales, reports, etc., as well as the results of formal testing or examination.

tests, a factor-analysis of jobs and tests is desirable, in order to reduce both the qualifications required and the traits to be tested to a small number of relatively independent key-qualities or "factors" (6, 9).

The problems discussed in this paper may arise at almost any of these stages; but it is chiefly in the course of validation and standardization that their urgency has been most strongly felt. In work for the British forces, the groups accessible for trying out the methods proposed, and the records available for validating those methods, are of a somewhat unusual nature. Consequently the commoner statistical procedures have often failed to yield satisfactory results; and certain further adaptations have proved indispensable. The devices I propose to describe were first given a practical trial in limited studies made by members of our Department during the early months of the war. Since 1941, as a member of the small committee of three, appointed by the Adjutant General to advise on problems of Personnel Selection, I have had data for Army recruits submitted for analysis on a far larger scale. Much the same problems have presented themselves in selection-work for the Navy and the Air Force. In these more recent studies the samples measured or assessed have often run into many thousands, so that probable errors shrink to small dimensions. As a result, where approximate formulas have been put forward, it has been possible to check the amount of inaccuracy involved with reasonable precision.

Three main questions constantly recur. (i) How are we to estimate the validity of the tests, when the only available criterion takes the form of a two- or threefold classification? (ii) How can we estimate their validity for the general population of recruits, when the data for the criterion refer only to selected and relatively homogeneous samples? (iii) Having finally decided upon our tests, how are we to determine the most appropriate borderline or pass-mark, in order to discriminate between those who are suitable and unsuitable for a given type of work?

Problems analogous to these faced the practical psychologist a quarter of a century ago, when educational and vocational psychology was still in an experimental phase. In pioneer work for an official body, like the London Educational Authority, the school psychologist encountered much the same difficulties as the military psychologist encounters today in carrying out research on personnel selection for the fighting services; and many of the methods adopted to overcome the earlier limitations might with advantage be revived. Many of the older devices lie buried in official reports; the newer have not yet appeared in print. Hence a brief summary of the procedures that

have proved most useful will perhaps be welcome, not only to military psychologists dealing with similar problems elsewhere, but also to research-workers in the new psychological fields that will call for scientific study when the war is over.

1. *Validation of Tests with Qualitative Criteria.* In order to assess the relative values of a series of different tests, it is essential to have independent gradings of the testees' efficiency at the tasks for which the tests have been designed. In educational research the requisite criterion can nowadays generally be procured from teachers, trained in psychology, and familiar with the conditions of statistical analysis; commonly it takes the form of reliable scores, standardized examination-marks, or trait-ratings conforming to some pre-arranged scale. But in vocational psychology, and particularly in work for the fighting services, detailed gradings of this kind are frequently unobtainable. Often the sole available criterion consists in a bare twofold or threefold classification, generally couched in qualitative terms —"Good," "Bad," "Indifferent," or "Suitable," "Unsuitable," "Doubtful."

In such cases, one simple expedient may be advocated at the outset. Wherever possible, reports on a threefold basis should be secured from two or more independent sources. Then, provided the reliability coefficient (as is usually the case) is not much above .70, by adding marks for *two* such reports for every man, we obtain totals that will form a reasonable distribution on a 5-fold scale; by adding marks for *three* reports, we obtain one on a 7-fold scale; and, with the criterion in this more finely differentiated form, the ordinary product-moment method can be employed, with Sheppard's adjustments to correct for coarseness of grouping if desired.

a. *The Tetrachoric Correlation and a Trigonometrical Approximation.* Where only a twofold grading is available for both test and criterion, a product-moment coefficient can still be calculated, but will not, without further transformation, yield coefficients comparable with those derived from graded variables having a normal distribution. The problem is one of the oldest in test-validation. When first standardizing the original Binet-Simon scale for use in British schools some thirty years ago, I considered it essential to start by assessing the validity, not merely of the scale as a whole, but also of each component test: this was, I believe, the earliest effort at what is now known as item-analysis (4, p. 205). Here, not only the criterion, but also the test-score ("pass" or "fail") was in a simple twofold form. The procedure suggested was either to compute Pearson's "tetrachoric" correlation, or (for rough provisional inquiries) to take the point-distribution correlation (ϕ), standardize it according to

Yule's method, and correct the coefficient by the familiar trigonometrical equation when comparison with a normal correlation was required. To save time and labour, abacs were drawn up both for the tetrachoric and for the standardized and unstandardized point-distribution coefficients (cf. 4, p. 219, Fig. 27). Similar graphs for the phi-coefficient have since been published by Guilford (11); and tetrachoric graphs, on a slightly different basis, by Thurstone and his colleagues. In my own graphs the two scales were arranged to give values (in percentages) for β/p and δ/q (in Kelley's notation, 5, p. 254) or $c/(c + e)$ and $f/(f + d)$ (in Thurstone's); in Thurstone's graphs they are arranged to give values for q' and δ. Thus, no matter what borderline be taken, my tetrachoric diagrams are all of the same size and shape—namely, square; and, for the present purpose at any rate, this would seem to render their usage at once easier and more accurate. The publication of the second volume of Pearson's Tables (2) has rendered the labour of computing such coefficients and constructing such graphs very much lighter.

When the tests are arranged to yield measurements on a continuous, graded scale, while the criterion is in twofold form, the so-called biserial correlation, being based on all the available information, should furnish the most accurate over-all estimate: nevertheless, where the frequency distribution is erratic, and where interest centres chiefly on test-validity near some particular borderline, the tetrachoric method still seems preferable. Indeed, provided the numbers are large and the assessments more or less standardized to a normal scale, Pearson's tetrachoric method (averaged if necessary for two or more dichotomies) provides so speedy and so close an approximation that we have regularly substituted it for the full product-moment method in most of our preliminary studies. For speedy item-analysis, where the criterion borderlines are *constant* for all items, we use a special graph which enables the investigator to read off values for tetrachoric r as soon as he knows what proportions of the highest and lowest groups, respectively, pass the item to be validated. Making the simplest possible assumptions, it is easy to show that the difference between two criterion groups of equal size (q) is most reliable when each group consists of a tail cut off from a normal distribution by a dividing-line whose deviation (x' say) is precisely one half of its mean ($\bar{x}_1$), i.e., when $2\,x' = \bar{x}_1 = z_1/q_1$: this gives $q_1 = 27.03$ per cent. With the more complex conditions obtaining in practical work, a slightly smaller proportion, say 25 per cent, would furnish a somewhat better result. But in educational research a five-point scale, with frequencies of 5, 25, 40, 25, 5 per cent, has proved exceedingly convenient; and for Army gradings, Cattell's scheme, with frequencies proportionate

to 1, 2, 4, 2, 1, has been adopted. Thus graphs appropriate to top and bottom groups of 30 per cent or 10 per cent have been most commonly required.

Where the main borderlines *vary*, the following formula yields a quick and convenient procedure for calculating the approximate tetrachoric correlation. In what Pearson calls the "tetrachoric functions," the essential factors (those containing h and k in Pearson's notation or x and x' in Kelley's), are simply the successive Hermite polynomials (parabolic cylinder functions) appearing in the development of the generalized (hypergeometric) frequency series. This implies the possibility of an approximation-formula based on trigonometrical functions. Thus, when x and $x' = 0$, Pearson's equation in its full form reduces to the expansion of $\sin^{-1} r$ in a power series: so that $r = \sin \phi'$ (where $\phi' = (ad - bc)/N^2HK$, in the notation of Pearson and Elderton, or $(\alpha\delta - \beta\gamma)/zz'$ in Kelley's notation.) Similarly when x and $x' \neq 0$, it will be found that, by taking analogous expansions for both $\sin \phi'$ and $\cos \phi'$, and introducing the necessary multipliers, the first few terms of the original equation can be expressed in the following form:

$$r_t = \sin \phi' - xx'(1 - \cos \phi) + \tfrac{1}{2}(x^2 + x'^2)(\sin \phi - \phi \cos \phi). \quad (1)$$

Short tables can be prepared for the two trigonometrical expressions within brackets; and with their aid a student or computing clerk can calculate 8 or 10 tetrachoric coefficients in the time required to calculate one when interpolating from Pearson's tables, with far less risk of error (**10**). Provided neither x nor r_t is very large, the approximation is quite close; when q is .10 or less and r_t is .70 or over, the error begins seriously to affect the second decimal place.

b. *A Coefficient for Triserial Correlation.* As already noted, the criterion is often supplied in threefold form, while the test-results are graded. Now if we can reasonably assume that the threefold classification rests on a variable that is continuously and normally distributed, then an extension of the proof of the biserial coefficient will yield a procedure that I have called "triserial correlation." The derivation of the formula is as follows.

Let x denote the actual score for the continuous variable (i.e., the test); let y denote the theoretical score for the trichotomous trait (i.e., the criterion); and let the subscripts 1, 2, and 3 denote the three subgroups—top, middle, and bottom, respectively—into which the whole group is divided by the trichotomous classification. Then, assuming the entire correlation-surface to be normal and the regressions linear,

$$r = \frac{(\bar{x}_1 - \bar{x}_3)/\sigma_x}{(\bar{y}_1 - \bar{y}_3)/\sigma_y}, \tag{2}$$

i.e., the correlation coefficient expresses the ratio of (1) the difference between the extreme subgroups, as determined by the test, to (2) the "true" difference between the same subgroups, as based on the criterion (a formulation which may often be used with advantage in explaining the notion of correlation to beginners). If q_1 and q_3 represent the proportions found in the two subgroups, and z_1 and z_3 the ordinates at the points of subdivision, then $\bar{y}_1 = z_1/q_1$, and $\bar{y}_3 = - z_3/q_3$. Hence

$$r = \frac{(\bar{x}_1 - \bar{x}_3)/\sigma_x}{z_1/q_1 + z_3/q_3}. \tag{3}$$

If the standard deviation, σ_x, is unknown, we can express it in terms of the standard deviation of either of the subgroups, s_1, say, by the equation

$$\sigma_x = s_1 \div \sqrt{\left\{ 1 - r^2 \left[\frac{z_1}{q_1} \left(\frac{z_1}{q_1} - x_1 \right) \right] \right\}}\,. \tag{4}$$

This yields, as a formula for a "triserial correlation,"

$$r = d \div \sqrt{\{ (z_1/q_1 + z_3/q_3)^2\, s_1{}^2 + z_1/q_1 (z_1/q_1 - x_1) d^2 \}}\,. \tag{5}$$

The expression on the right has several useful variants which can be employed for special cases—e.g., for those in which the two standard deviations are widely different, or the numbers in the extreme subgroups are approximately equal.

c. *The Biserial Correlation with a Point-Distribution.* In many of our investigations the available information is more limited still: not only have the tests been applied solely to men representing the two extreme subgroups (the "outstandingly efficient" and the "definitely inefficient"), but there may be nothing to show what proportion the number in either subgroup bears to the total number in the complete sample from which they have been drawn. In practice the proportions are usually small, since the prime object has been to obtain a well-marked contrast. Moreover, each subgroup is generally truncated on both sides—on the distal side as well as on the proximal: directly or indirectly, a preliminary sifting will probably have removed the physically or mentally defective from the poorer section, and the most highly endowed from the better section. Under such conditions the best available procedure would seem to be what may be called the method of "biserial correlation for a point-distribution."

Let us derive the formula for a single test first of all. The simplest proof starts from the assumption that the individuals in the extreme subgroups vary so little among themselves that their two frequencies can be considered as concentrated each at a single point. However, as in most other cases, the correlation so deduced may also be interpreted in terms of the least squares principle, and consequently rendered independent of any special hypothesis about the way in which the variates are distributed. In the most general case of all, the two subgroups may be regarded as consisting respectively of individuals who possess, and individuals who do not possess, some distinguishing quality.

Let the total number be N; and of these let N_1 and N_0 be the number of persons with and without the quality in question $(N_1 + N_0 = N)$. Then, if we assign to each person a mark of either 1 or 0 according as the distinguishing quality is present or not, the means of the two subgroups will be $-N_0/N$ and $+N_1/N$, respectively, and the standard deviation of the entire group will be $\sqrt{N_1 N_0}/N$. Now let x denote the measurements obtained with the test in question, and $\bar{x}_1$ and $\bar{x}_0$ the means of such measurements for each of the two subgroups. Substituting these results in the ordinary formula for regression, we obtain almost at once

$$r_p = \frac{(\bar{x}_1 - \bar{x}_0)\sqrt{N_1 N_0}}{\sigma_x (N_1 + N_0)} = \frac{d\sqrt{N_1 N_0}}{\sigma_x N}, \tag{6}$$

where d denotes the difference between the two means and r_p the correlation between the test and the dichotomous criterion.

When we are dealing, not with a single test only, but with a battery of tests, we want to assign to each test its own appropriate weight. The weights will, of course, be proportional to the partial regressions. If then the criterion-correlations are based on an assumed point-distribution, it follows that the values of the partial regressions can be directly computed by the simple expression:

$$\mathbf{w}' = \mathbf{r}' R^{-1} = k\,\mathbf{d}' R^{-1}, \tag{7}$$

where w denotes the row-vector of partial regressions, r that of the correlations with the criterion, d that of the differences in standard measure, R (as usual) the matrix of intercorrelations, and k (a scalar) the constant term $\sqrt{N_1 N_0}/\sqrt{N_1 + N_0}$ (which can be omitted if proportional values alone are sufficient): so that the criterion correlations need not be calculated at all.

The formula may also be employed when all the men in a given batch have to be allocated to one or other of two alternative occupa-

tions. For example, in the Air Force, psychologists have attempted to discriminate between men who would make good bombers and men who would make good pilots; and it would naturally be convenient if this could be done by using the same set of tests for both allocations. Accordingly, we administer a mixed battery of tests to two suitably selected subgroups, one consisting of highly efficient bombers and the other of highly efficient pilots: then the regression coefficients calculated by the foregoing formula will indicate both the relative merits, and the most suitable weightings, for each of the tests employed, when we are considering the assignment of the man to one type of duty rather than to the other. In more general inquiries, wherever we have to deal with categories based on the simple presence or absence of an ungraded characteristic (e.g., sex or a Mendelian unit-trait), this formula would seem to embody the most suitable procedure.

Where the same tests have been applied to complete batches of men, who were given not two, but a number of alternative training recommendations, and where subsequent records are available stating the efficiency of each in the work to which he has been allocated, the most satisfactory and most rigorous procedure for determining the validity of the various tests for the alternative recommendations is provided by the analysis of variance and covariance. Hitherto psychologists have been somewhat chary of adopting this newer technique, and have kept mainly to the older formulas for multiple correlation. In our experience each mode of approach has its own particular uses; and in another paper I have endeavoured to set out more fully the suitability and the limitations of the two alternative techniques for psychological problems of various kinds (7).

2. *The Influence of Selection.* In validating tests for the Army, as in most other vocational investigations, it is rare to obtain multiple correlations much above .5. So far, with few exceptions, it is only for duties, like those of clerks, signallers, or engineers, for which both the qualifications and the appropriate tests are of a semi-scholastic type, that higher coefficients have been found. The lowness of the figures would seem to be due to much the same factors as have handicapped the vocational investigator in ordinary industrial work. First, it is exceedingly difficult to procure good independent criteria. Too often the estimates of the men's success at the jobs to which they have been assigned are based on each man's general efficiency and character rather than on his specific suitability for the duties in question. Further, assessors in different units, and even assessors in the same unit reporting on men of different efficiency, give different weight to different qualities: with the poorer type, it is sheer lack of intelligence that seems most commonly noticed; with the better type, qual-

ities of character or leadership. Secondly, as is shown when two or more independent ratings are obtainable for the same batch of men, the "reliability" of the assessments is exceedingly low, particularly where the batches are large. In a few small groups with which I have worked, where the officers assessing the men have taken special care over their gradings and have produced fairly reliable assessments, the multiple correlation rises above the figures obtained elsewhere by 20 or 30 per cent; in such cases they may reach a value of .65 or more—results that approximate to those commonly found in educational inquiries where conditions of assessment are more satisfactory. Thirdly, under service conditions of examination and interview, the tests and personality ratings are themselves less reliable than might be desired. With the Kuder-Richardson full formula the reliability usually lies between .80 and .95 for verbal and arithmetical tests, and between .55 and .75 for non-verbal or "performance" tests: for estimates of character traits it is nearly always well below .60. Finally, the data available for validating the tests relate, almost inevitably, not to the entire population of candidates or recruits, but to selected groups of men who have already been picked out for their supposed ability in the very tasks which the tests are assessing. As a result, the groups on which the tests are validated are bound to be more homogeneous in the traits concerned than the general mass of recruits from which those groups have been drawn. Consequently the correlations obtained must be appreciably reduced.

This factor of selection is rarely considered; yet it is one of the most important, and can readily be allowed for. Attenuations arising from this source have been repeatedly pointed out in early studies with educational tests (e.g., 3, pp. 52, 62-3; 4, pp. 205-6); yet teachers and educational psychologists, in reporting their researches, still record with disappointed surprise the low correlations they discover on applying mental or scholastic tests to the relatively homogeneous groups of children with whom they usually work (e.g., mental defectives in a special school, successful scholarship candidates followed up at secondary schools, or, most frequently of all, the pupils of a single school class, instead of the complete and heterogeneous age-group—all highly selected samples). Similarly, in researches on vocational guidance, the small correlations obtained from the selected groups employed have continually been cited as evidence for low predictive value of psychological tests, or for the absence of any appreciable "general factor," in apparent ignorance of the true causation (9). Some statistical device, therefore, is urgently needed whereby we can correct for the disturbance introduced by unavoidable selection.

The obvious method is the inverse of that devised by Karl Pear-

son, over forty years ago, for forecasting the probable effects of natural selection on the physical characteristics of an evolving species (1). By inverting his procedure, we are accepting the familiar risks of extrapolation. But, before adapting it for educational inquiries, I attempted several empirical checks by taking instances where the values for the unselected population could be directly ascertained (12 and refs.): it seemed clear that, provided the conditions assumed by the formulas were not flagrantly violated, the corrected values generally provide far better estimates of the true values than the uncorrected, and could always be used to verify the absence of serious distortion.

Pearson's proof, and the arithmetical work which his formulas entail, are, as he himself declares, "somewhat formidable." The expressions he obtained were originally derived by postulating that, both in the selected and in the unselected groups, the frequency-distributions "follow the normal or generalized Gaussian law." Wider applicability, however, is attainable if we proceed from what are now the commoner assumptions in correlational work—namely, that any regression we are dealing with may be adequately represented by a linear function and that the variances of the different arrays are approximately equal: (if desired, these assumptions can be formally tested for any given set of data by undertaking an analysis of the variance and applying the so-called L_1-test of Neyman and E. S. Pearson). On this broader basis, we can reach a far simpler proof, and secure convenient formulas which are much easier to apply in actual practice.

Accordingly, let a denote those tests for which we have complete information, i.e., the tests which have been given both to the total or unselected population and to the selected sample drawn from it: in the simplest cases these tests will be those on which the selection has actually been based; we may therefore term them the "selective variables." Let x denote tests which have been applied to only *one* of these two groups: these we may term the "non-selective variables." In Pearson's problems the group for which all the constants are given was always the total population, or at least an unselected sample of it; in our problems it is usually the smaller and relatively homogeneous subgroup. Further, let S denote the matrix of standard deviations; R, the correlations (with unity in the diagonal); C, the covariances (with variances in the diagonal); W, the partial regressions for raw scores; and B, the partial regressions for standardized scores ("beta-coefficients"). Following the common convention, capital letters will indicate constants for the larger, unselected group, and the corresponding small letters those for the smaller, selected group. The sym-

bol $H_x = \Sigma_x/\sigma_x$, or in matrix notation $H = Ss^{-1} = h^{-1}$, will be used to denote the proportionate changes in variability produced in the several tests by reversing the process of selection; and $R_{x(a)}$ to denote the multiple correlation between any non-selective variable (x) and the best linear combination of the selective variables (a).

Now, with the conditions under which the smaller group is supposed to have been selected, the selection itself cannot (except for errors of random sampling) affect the regressions for *raw* scores: that is, $W'_{xa} = w'_{xa}$. Hence $S_x B'_{xa} S_a^{-1} = W'_{xa} = w'_{xa} = s_x b'_{xa} s_a^{-1}$; and

$$B'_{xa} = H_x^{-1} b'_{xa}\, h_a^{-1} = H_x^{-1}\, U'_{xa} \text{ (say)}. \tag{8}$$

It follows that

$$R'_{xa} = B'_{xa}\, R_{aa} = H_x^{-1}\, U'_{xa}\, R_{aa}, \tag{9}$$

and

$$R^2_{x(a)} = B'_{xa}\, R_{ax} = H_x^{-1}\, (U'_{xa}\, R_{aa}\, U_{ax})\, H_x^{-1}. \tag{10}$$

For purposes of practical assessment, we merely require U'_{xa}, the *proportionate* values for the beta-coefficients. For theoretical purposes, however (e.g., calculating the corrected multiple correlation by equation 10), we must also find H_x, though we need not explicitly calculate either the corrected regressions (B'_{xa}) or the corrected correlations with the criterion (R'_{xa}), unless we wish to study those particular constants in and for themselves. If we eliminate all influence of the selective variables from both the unselected and the selected groups, the residual variances must be equal: thus

$$\Sigma^2_x(1 - R^2_{x(a)}) = \Sigma^2_{x.a} = \sigma^2_{x.a} = \sigma^2_x(1 - r^2_{x(a)}). \text{ Hence}$$

$$H^2_x = 1 - r^2_{x(a)} + H^2_x\, R^2_{x(a)} = 1 - b'_{xa}\, r_{ax} + U'_{xa}\, R_{aa}\, U_{ax}. \tag{11}$$

If covariances have already been calculated, we may take

$$U'_{xa}\, R_{aa}\, U_{ax} = b'_{xa} s_a^{-1}\, C_{aa}\, s_a^{-1}\, b_{ax}.$$

If we prefer to work throughout with covariances rather than correlations, the same premises lead with equal ease to the more general equations:

$$C'_{xa} = w'_{xa}\, C_{aa} = c'_{xa}\, c_{aa}^{-1}\, C_{aa} \text{ and } C_{xx} = c_{xx} + w'_{x'a}\, (C_{ax} - c_{ax}).$$

In matrix form the foregoing formulas are available primarily for estimating the effects of multivariate selection. If the selection is univariate, they reduce to simpler expressions, which can, in point of fact, be independently deduced from the ordinary formulas for partial correlation. When selection is complete, and the standard deviation for the selected variable consequently zero, the formulas finally resolve into the well-known equations for partial correlation. They

may, indeed, be regarded as generalized versions of the latter. In the familiar formulas for partial correlation we assume that we select *at* a given value; in the more general selection formulas, we assume that we select *about* a given value.

The application of these corrections has proved exceedingly instructive. With the test-batteries hitherto employed, it would seem, the multiple correlations obtained from selected groups require as a rule to be raised by 10 to 15 per cent to indicate the validity of the tests when applied to an unselected population. Thus, where the Army scheme for personnel selection is working satisfactorily, and reliable data are available, the validity coefficients among recruits before selection may be very roughly estimated as reaching figures in the neighbourhood of .70 or .75. Their efficiency, therefore, accords reasonably well with that usually reported in pre-war validation-studies in both the educational and the vocational field.

The improvement in efficient selection achieved by means of such tests may be gauged from figures secured in following up the men who have taken up the more important army trades; (it is in these branches that a man's eventual failure or success can be most readily assessed). Broadly speaking it would appear that the procedure now adopted has reduced the number of failures to about one-half of that observed among men who had been selected by the older procedures operating in the Army, and to about one-third of that observed among men selected for such jobs by the Ministry of Labour. No doubt, when the time arrives, those who have been officially in charge of the work will be able to give precise and detailed figures for every branch. (Cf. 7, 10.)

A word of caution is necessary. In work with partial correlations it is often forgotten that the formulas may legitimately be applied only under certain conditions: if the variable eliminated includes unsuspected factors which do not all enter into the other variables from which we are proposing to eliminate it, or if it is itself wholly or in part dependent on those variables, or again if it depends on unsuspected common factors which affect all the variables in very different degrees, then the ordinary formula is no longer applicable: it would, in fact, "partial out" too much. The same is true of the selection formulas; and if the assumptions implied in their proof are not adequately fulfilled, then the figures deduced by their aid are likely to exaggerate the influence of selection. Accordingly, besides knowing the formal expressions required to solve the general problem, we must also know something about the material nature of the particular data for which we are proposing to use them: we must, in short, possess some justifiable hypothesis about the structure of the causal

framework that links together the different variables with which we are working.

3. *Determining Borderlines.* After we have discovered what tests are most suitable for our purpose, and after we have determined their validity and weight, both individually and combined in a battery, one problem still remains, namely, to fix the borderline or pass-mark by which we decide whether any given candidate is to be rejected or selected. On theoretical grounds, though not always on practical, the best pass-mark would manifestly be that which minimizes the discrepancies between the findings of the test and the findings of the criterion. On this assumption suitable formulas are not far to seek, for in educational psychology the problem has long been familiar (3, 4).

In early work in the London schools one of the first tasks of the Education Authority's psychologist was to determine, in terms of test-measurements, a general borderline for certifying mentally deficient children for transference to the "special schools." In their verdicts on *individual* pupils, both teachers and school medical officers not infrequently make serious errors: thus, in a general survey of the school population, it was found that the brightest 30 per cent of the children sent to special schools as certified defectives were actually brighter than the dullest 3 per cent who had been left in the ordinary schools as normal. On the other hand, it was agreed that their *general* notions of what type of child could or could not be taught in the ordinary elementary school formed the best available criterion. Accordingly, in formulating a theoretical borderline in terms of intelligence-tests, the principle to adopt seemed plain: namely, to draw the test-borderline at a point which would minimize the discrepancies between the test-results, on the one hand, and the results of the independent assessments of the teachers and doctors, on the other. The teachers were able to assess the individual children on a continuous scale of marks: the doctors merely classified the children into two groups—certifiable and uncertifiable. This entailed devising two alternative formulas. With this particular problem, both (as it happened) gave virtually the same result.

(i) Where the independent assessments, forming the criterion, are on a continuously graded scale, it is possible to calculate the full product-moment correlation (r) between the test (x, say) and the criterion (y, say). Then the regression equation for estimating what criterion-value will correspond to any particular test-mark will be $y = b_{yx}x = r \frac{\sigma_y}{\sigma_x} x$. With standard measure this reduces to $y = rx$;

and the equation to the correlation-surface can be written

$$z = \frac{1}{2\pi\sqrt{(1-r^2)}} e^{-\frac{1}{2(1-r^2)}(x^2+y^2-2rxy)} = \phi(x,y).$$

Now let k denote the known value of y which marks the borderline between those who, on the basis of the criterion (y), would be selected and those who would not; and let h denote the required value of x which will serve as the best borderline in terms of the test—the "pass-mark," as it may be called. The discrepant cases whose number we seek to minimize will be made up of two groups: (i) those who pass on the criterion and fail in the test and (ii) those who pass in the test and fail on the criterion. Their proportionate numbers will be given by $\int_k^\infty \int_{-\infty}^h \phi(x,y)\,dx\;dy$ and $\int_{-\infty}^k \int_h^\infty \phi(x,y)\,dx\;dy$, respectively. If h is to be so chosen that the sum of these two integrals is to be a minimum, then the first partial differential of that sum, taken with respect to h, must be zero, and the second must be negative.

On differentiating we have for the first differential $\int_k^\infty \phi(h,y)\,dy - \int_{-\infty}^k \phi(h,y)\,dy$. Now let us convert deviations about the horizontal axis to terms of deviations about the regression line $y = rh$. Changing the variable accordingly to $y' = y - rh$, and writing $k' = \dfrac{k - rh}{\sqrt{1-r^2}}$ for the new limit, we have, as the requisite condition, $\int_{k'}^\infty e^{-\frac{1}{2}y'^2}\,dy' - \int_{-\infty}^{k'} e^{-\frac{1}{2}y'^2}\,dy' = 0$. But this will be attained only when $k' = 0$. In that case $\dfrac{k - rh}{\sqrt{1-r^2}} = 0$ and $k = rh$, or, when the deviations have not been converted into standard measure,

$$\frac{k}{\sigma_y} = r\frac{h}{\sigma_x}, \text{ that is, } h = k/b_{yx}. \qquad (12)$$

Thus, the "minimal discrepancy pass-mark" is found by dividing the value of the given criterion borderline by the regression of the criterion on the test. We may render this simple result more intelligible by expressing the rule it yields as follows: take for the required test-borderline that particular pass-mark from which you could deduce the given criterion-borderline, if you endeavoured to predict the latter from the former in the usual way by means of the appropriate

regression equation. Thus (to round off the actual figures in my *Report* by way of illustration) suppose that the teachers' borderline is drawn at a point equivalent to an I.Q. of .76, i.e., .24 below the average, and that the correlation between the teachers' educational ratings and the I.Q. as obtained with the Binet tests is .80: then the minimal discrepancy borderline for defectives should be drawn at an I.Q. of .70, i.e., .30 below the average, since $.80 \times -.30 = -.24$. When h has been thus determined, the regression line, and therefore the criterion borderline, will cut the "array" of type h into two equal halves. Any child who obtained an I.Q. of precisely .70 stands equal chances of being nominated or not nominated as fit for a special school by a teacher who knows nothing of the test-result (4; cf. 7, "Note on Wastage Coefficients").

(ii) Where the independent assessments consist of a simple twofold classification, an alternative formula was used. It was derived as follows.

Let N_1 and N_2 be the numbers in the two groups, and x_1 and $-x_2$ the distances of the point of intersection of the two frequency-curves from the central ordinate of each; and let $d = (x_1 + x_2)$ be the distance between the two medians. Assuming that the curves are both approximately normal, the ordinate at x_2, being common to both curves, will be $\frac{N_1}{\sigma_1\sqrt{2\pi}}\, e^{-\frac{x_1^2}{2\sigma_1^2}} = \frac{N_2}{\sigma_2\sqrt{2\pi}}\, e^{-\frac{x_2^2}{2\sigma_2^2}}$. Taking logarithms of both sides and substituting for x_1, we have

$$\log_e \frac{N_1}{\sigma_1} - \frac{(d - x_2)^2}{2\,\sigma_1^2} = \log_e \frac{N_2}{\sigma_2} - \frac{x_2^2}{\sigma_2^2}.$$

Hence

$$x_2 = \frac{\sigma_1^2\, d - \sigma_1\,\sigma_2 \sqrt{\left\{d^2 + 2(\sigma_1^2 - \sigma_2^2)\,\log_e \frac{\sigma_1 N_2}{\sigma_2 N_1}\right\}}}{\sigma_1^2 - \sigma_2^2}. \qquad (13)$$

If the curves are appreciably asymmetrical, the calculation of σ_1 and σ_2 should be based on the intersecting halves only, i.e., on the upper half of the lower curve and the lower half of the upper curve. This formula seems preferable where the individual test-assessments are likely to be more reliable than the classification of the individuals according to the criterion.

Both formulas were originally put forward for computing tentative borderlines for the transference of pupils to trade schools, secondary schools, and schools for the mentally defective (4). They have since demonstrated their value in theoretical discussions of analogous

problems arising in vocational selection for industry. But for administrative requirements they are not necessarily the most suitable. With graded variables, to minimize discrepancies as such is not of the first importance. Unless the correlation is abnormally low, the majority of the discrepancies are likely to be comparatively slight; and in practice the psychologist is seldom free to suggest whatever borderlines he likes.

In the early stages of personnel selection for the Army, when a relatively simple scheme had to be contemplated, it was helpful to compute figures for exclusion levels and for wastage coefficients by formulas like the foregoing (7). But, as the organization has grown more complex, and greater experience gained, administrative considerations—such as the current state of demand and supply, and the need for allocating reasonable proportions of efficient and less efficient men to all the main branches—have tended more and more to determine the borderlines laid down. Consequently, the work of the statistical psychologist has of late been directed rather to the construction of norms and minimum "profiles" for each important type of job in terms of the tests most relevant to it.

When the war is over, British psychologists will once again be called in by the Civil Service Commission to supply tests for ex-service men and others, similar to those constructed by Prof. Spearman and myself during the years that followed the last demobilization. Research has already started on the production of suitable tests; and the methods described in the foregoing pages will, it is believed, greatly facilitate the preliminary validations. Some of the formulas may have an even wider application.

REFERENCES

1. Pearson, K. On the influence of natural selection on the correlation and variability of organs. *Phil. Trans. Roy. Soc.* (A), 1902, **200**, 1-66.
2. ———. Tables for statisticians and biometricians, Pts. I and II. London: Biometric Laboratory, University College. Pt. I, 1914; Pt. II, 1931.
3. Burt, C. The distribution and relations of educational abilities. London: P. S. King, 1917.
4. ———. Mental and scholastic tests. London: P. S. King, 1921.
5. Kelley, T. L. Statistical method. New York: Macmillan Co., 1923.
6. Burt, C. Principles of vocational guidance. *Brit. J. Psychol.*, 1924, **14**, 336-352.
7. ———. Memoranda on psychological testing in the Army. 1941-4 (circulated privately).
8. ———. Psychology in war. (Presidential Address, British Psychological Society, 1942.) Reprinted (abridged) in *Occupational Psychology*, 1942, **16**, 95-111.

9. ———. The use of statistics in vocational psychology. *Occupational Psychology*, 1942, **16**, 164-174.
10. ———. Military psychology in Great Britain. (Presidential Address, British Psychological Society, 1943.)
11. Guilford, J. P. The phi-coefficient and chi square as indices of item validity. *Psychometrika*, 1941, **6**, 18.
12. Burt, C. Validating tests for personnel selection. *Brit. J. Psychol.*, 1943, **34**, 1-19.

PSYCHOMETRIKA—VOL. 9, NO. 4
DECEMBER, 1944

FACTOR PATTERN OF TEST ITEMS AND TESTS AS A FUNCTION OF THE CORRELATION COEFFICIENT: CONTENT, DIFFICULTY, AND CONSTANT ERROR FACTORS

ROBERT J. WHERRY AND RICHARD H. GAYLORD
UNIVERSITY OF NORTH CAROLINA

A dilemma was created for factor analysts by Ferguson (*Psychometrika*, 1941, **6**, 323-329) when he demonstrated that test items or sub-tests of varying difficulty will yield a correlation matrix of rank greater than 1, even though the material from which the items or sub-tests are drawn is homogeneous, although homogeneity of such material had been defined operationally by factor analysts as having a correlation matrix of rank 1. This dilemma has been resolved as a case of ambiguity, which lay in (1) failure to specify whether homogeneity was to apply to content, difficulty, or both, and (2) failure to state explicitly the kind of correlation to be used in obtaining the matrix. It is demonstrated that (1) if the material is homogeneous in both respects, the type of coefficient is immaterial, but (2) if content is homogeneous but difficulty is not, the homogeneity of the content can be demonstrated only by using the tetrachoric correlation coefficient in deriving the matrix; and that the use of the phi-coefficient (Pearsonian r) will disclose only the non-homogeneity of the difficulty and lead to a series of *constant error* factors as contrasted with *content factors*. Since varying difficulty of items (and possibly of sub-tests) is desirable as well as practically unavoidable, it is recommended that all factor analysis problems be carried out with tetrachoric correlations. While no one would want to obtain the constant error factors by factor analysis (difficulty being more easily obtained by counting passes), their importance for test construction is pointed out.

Introduction

Fergeson (1) has shown empirically that test items of varying difficulty, when correlated according to the phi coefficient formula

$$r_{ij} = \frac{p_{ij} - p_i p_j}{\sqrt{p_i p_j q_i q_j}}, \tag{1}$$

yield a matrix whose rank is greater than 1 (the rank actually equals the number of difficulty levels), even though the test from which the items are drawn be homogeneous with respect to content. Thus he creates an apparent dilemma. According to Thurstone (and other factor theorists) the operational test of homogeneity among a set of tests (or items) is that their correlation matrix have a rank of 1. Yet Fergeson appears to have demonstrated the impossibility of verifica-

tion by means of the operational definition! But this must be scientific double-talk. Three possibilities exist: (a) Fergeson's imputedly homogeneous test is not homogeneous, (b) Fergeson has not correctly applied the operational test, or (c) the term homogeneous, and consequently its indicated operational test, is an ambiguous matter. We shall show that the third alternative is correct, and that therefore Fergeson was guilty of either (a) or (b), depending upon how one views the problem. Since the whole matter is a basic one for test analysis it seems worth while to investigate it further.

Factors Based on Content versus Factors Based on Difficulty of Test Items

Let us examine the characteristics of the simplest scale which generally would be admitted to be homogeneous with respect to content. Such a scale would be one measuring length in the case of a population of line segments possessing the single dimension of length. The measuring device is a straight stick which has a number of calibrations marked along its edge. These calibrations are numbered from one end with numbers ranging from 1 to n. The calibrations are not necessarily equidistant since that criterion of measurement is seldom satisfied in psychological tests or scales. As each line segment is placed alongside the scale with one end placed in juxtaposition with the zero end of the scale we could make a single judgment as to the last calibration exceeded by the other end, or we could make successive judgments as to which of the calibrations had been exceeded and then summate these, the number or score being the same in either case. The second alternative is identical with counting the items passed on a homogeneous psychological test. For the calibrations (items) and for the judged lengths of the segments (scores) we then have the following characteristics:

1. The items (calibrations) are arranged in ascending order of scale value $X_1 < X_2 < X_3 < X_4 \cdots < X_n$ where X is the scale value of any item and the subscript is the ordinal number assigned to it.

2. The number of passers for each item will be inversely proportional to the ordinal numbers assigned—but will not necessarily correspond exactly, since no assumption was made as to the nature of the distribution of line segments (the population), thus

$$P_1 \geqslant P_2 \geqslant P_3 \geqslant P_4 \geqslant \cdots \geqslant P_n ,$$

(The possibility that two calibrations might be coexistent or that no members of the population will have sizes

between some of the adjacent calibrations, i.e., that two test items might possess the same difficulty, is taken care of by using = as well as > at this point.)

where P is the number of passers and the subscript is the ordinal number of the item or calibration.

3. The number of individuals passing both of any pair of items will always be equal to the number passing the item of higher ordinal number, thus

$$P_{s1} = P_1 \leqslant P_s \, , \tag{2}$$

where P_{s1} is the number of passers the two items have in common, P_s is the number passing the item with the smaller ordinal number, and P_1 is the number passing the item with the larger ordinal number (the harder item).

4. The correlation between any two items will then be:
(a) Using equation (1) proposed by Fergeson, and substituting the value of P_{s1} from (2) above,

$$\begin{aligned} r_{s1} &= \frac{p_{s1} - p_s p_1}{\sqrt{p_1 p_s q_1 q_s}} = \frac{p_1 - p_s p_1}{\sqrt{p_1 p_s q_1 q_s}} \\ &= \sqrt{\frac{p_1}{p_s}} \sqrt{\frac{1 - p_s}{1 - p_1}} = \sqrt{1 - \frac{(p_s - p_1)}{p_s(1 - p_1)}} \, ; \end{aligned} \tag{3}$$

i.e., r_{s1} will equal unity only when $p_s = p_1$, and will be smaller than unity by an amount proportional to the distance between s and 1 on the scale, the greater the difference in difficulty the lower being the correlation.

(b) If the tetrachoric formula is used the correlation will be 1.00 in every case regardless of difficulty differences.

We have exposed and explained in the above the entire dilemma supposed to exist on the basis of Fergeson's discovery:

(a) If we assume as we have above that our scale is homogeneous *with respect to content* (Fergeson's homogeneous test is equivalent to this scale), then it follows that the proper operational test of this condition is the use of the tetrachoric equation for the obtaining of the intercorrelations of items along with a determination of the rank of the matrix by factor analysis. In the special case above where the test was not only homogeneous but also perfectly reliable all inter-r's will be 1.00 and the factor analysis will correctly indicate the rank of 1 for the matrix. Thus it would tell us that all calibrations did indeed belong to a single scale or stick.

(b) If we apply instead the phi coefficient as Fergeson did, it follows from formula (3) that the size of the coefficients obtained will be contingent upon difficulty. Of course we could proceed as Fergeson did to extract the factors showing that although the test was homogeneous with respect to content it was not so with respect to difficulty and finding out as well which items were near each other in difficulty. In the usual test situation, however, such factors would be poor measures of difficulty if content varied and poor measures of content if difficulty varied. It would be much simpler to determine difficulty by counting passers, and to determine homogeneity of content by analysis of the matrix of tetrachoric coefficients.

We may now return to the three possibilities mentioned in the introduction. We see that Fergeson's test was homogeneous with respect to content but not so with respect to difficulty. As usually stated the operational test for homogeneity fails to take account of this ambiguity. We see now that if it is homogeneity of content that is to be tested (the usual practical situation), we must use the tetrachoric correlation coefficient which is not contingent upon difficulty. Of course, homogeneity of content having already been established, we can test for homogeneity of difficulty (the Fergeson situation) by using the phi coefficient which is contingent upon difficulty, although difficulty is more easily determined otherwise.

It should be clear now that the *difficulty factors* obtained by Fergeson were the result of the impact of item difficulty upon a coefficient which is sensitive to difficulty. We hope that the insecurity which must have been felt by factor analysts dealing with test items since the publication of the Fergeson article will be dispelled in those cases where the tetrachoric coefficient has been used.

Role of Constant Error Factors among Items in Test Construction

While we would not emulate Fergeson in his mode of determining difficulty levels, or factors based on difficulty, we would like to point out their *theoretical* meaning and their *practical* importance. The factors obtained by Fergeson were not mere mathematical artifacts even though they were logical fallacies. Let us consider what an easy item j in our previous scale accomplishes as the measurer of our population of line segments. Item j (see Figure 1) would tell us that persons A, B, and C failed while D, E, F, $\cdots$, L, M, and N passed the item. Now persons A, B, and C being lumped descriptively as "less than j" would be well described with *small constant errors,* while persons D, E, F, $\cdots$, L, M, and N lumped descriptively as "greater than j" would be very poorly described with widely varying but on the average *large* constant errors. Conversely item k in the figure would

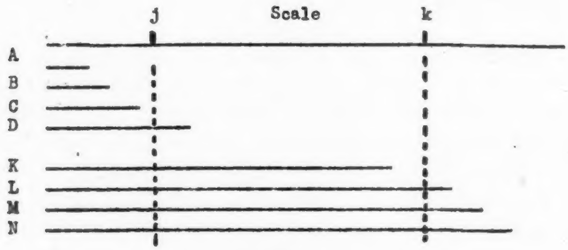


FIGURE 1
Scale and Line Segments

lump together persons A, B, C, D, ···, J, and K as "less than *k*," introducing *large* constant errors for persons A, B, and C who were adequately described by *j*, while item *k* would lump together only persons L, M, and N as "greater than *k*" and thus describe them with *small* constant error, whereas they had been badly classified by item *j*. We see that theoretically the *difficulty* factors of Fergeson are allied to differential determination of the degree of *constant error* for various parts of the population.

This has been previously shown by Richardson (3) and Walker (4) to be an important determiner if one wishes to adequately select an extreme portion of a population of individuals. If one wishes to eliminate only a small group of the poorest applicants a test composed of only easy items (homogeneous with respect to difficulty at say the *j* level) should be employed, since such a test will measure such individuals as A, B, and C most accurately with the *least* possible constant error. Similarly if one wished to pick only a small part of the population at the upper end of the distribution, items homogeneous with respect to difficulty at the *k* level would be employed. On the other hand if one wishes to accurately rank *all* persons along the entire scale items of *all levels* of difficulty are needed. A test composed only of items of medium difficulty could tell us accurately only to which half of a distribution, upper or lower, a person belonged, and everybody would be moderately inaccurately placed, i.e., have moderate constant error in his score. These theoretical conclusions have been empirically verified in the papers of Walker and Richardson cited above. One would not therefore want to follow Fergeson's suggestion of using items of only one difficulty level in establishing the homogeneity of content of his test (although as Fergeson pointed out this would make the phi coefficient applicable) if he wanted a scale which would accurately rank an entire population.

Factors based on Content versus Factors based on Difficulty among Tests

The importance of the correlation coefficient to be used extends to factorization of tests as well as to items. Consider a test composed of n items of varying difficulty but still obeying the rule that $P_{1s} = P_1$. If we assume that our test is divided into two halves so that T_I is composed of items a, b, c, $\cdots$, $n/2$ and T_{II} is composed of items a', b', c', $\cdots$, $n'/2$, where $P_a = P_{a'}$, $P_b = P_{b'}$, $\cdots p_{n/2} = P_{n'/2}$, then the scatter diagram for scores on the two halves would be shown in Figure 2. The

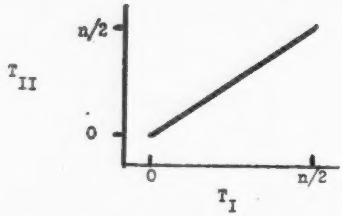


FIGURE 2

Scatter diagram showing distribution of scores on test halves equated for difficulty

Pearsonian product moment correlation coefficient would be 1.00. If however we had our test divided into two subtests T_E composed of items a, b, c, $\cdots$, $n/2$ and T_H composed of items $n/2+1$, $n/2+2, \cdots$, n, where $P_a > P_b > P_c > \cdots > P_{n/2} > P_{n/2+1} > \cdots > P_n$, then the scatter diagram for the scores on the two halves would be as in Figure 3, since

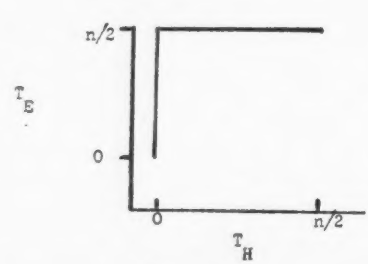
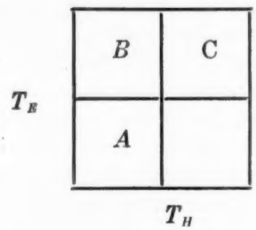


FIGURE 3

Scatter diagram showing distribution of test half scores when difficulty difference is maximized

only those persons making a zero score on the hard test could have lower than a maximum score on the easy test. The Pearsonian correlation would not be 1.00.

It would seem futile to apply a linear coefficient to this kind of scatter diagram, but Fergeson went to some pains to do just that. He selected a set of six subtests from a long supposedly homogeneous test, so that the average difficulties were .666, .524, .423, .316, .218, and .106, respectively. He then computed Pearsonian correlations between the test scores and demonstrated that the test matrix did not have a rank of 1.00, a conclusion which would follow at once from the two diagrams above. However, had Fergeson dichotomized his sub-tests and then used the tetrachoric equation* for intercorrelations the result would have been different. Even in our second scatter diagram above we would achieve:



and the tetrachoric r would be 1.00. Of course Fergeson was working with fallible test scores rather than perfect tests, so that even granting that his tests were homogeneous in content the r's would not necessarily be 1.00, but we would predict that the inter-r's would at least no longer be a function of difficulty, i.e., would yield *content* rather than *constant error* factors if the rank were not 1. Thus we see again that the Fergeson difficulty factors arose through a methodological error. While the test for homogeneity as commonly stated by the factor theorists fails to state which correlation coefficient should be used, it certainly did not imply that erroneously determined coefficients are permissible. Analysis of a matrix of linear coefficients obtained from skewed data seems unwarranted.

Of course *content* could shift with level of *difficulty* or difficulty shift because of a change in content. This is well illustrated in a study by Guilford (2) who applied factor analysis to tests (of 10 items each) of various levels of difficulty of the *supposedly* homogeneous function of pitch discrimination. This study is sometimes cited erroneously as an example of the Fergeson dilemma, but such is not

* One critic objected that this would be "violating quite violently" the assumption of normality upon which the tetrachoric is based since the distributions of test scores are *J*-shaped. This is not a valid objection, however, since it is assumed that the trait (not the scores) is normally distributed. True, the trait *may not* be normally distributed, but even if it is the scores will still yield the *J*-distribution in a dichotomy.

the case. Guilford did use dichotomization and tetrachoric inter-r's so that his resulting *difficulty* factors are actually a disproof of the homogeneity claimed for pitch discrimination (at varying levels of difficulty) rather than a mere verification of the fact that difficulty levels do lead to varying *constant errors*.

Since real content factors, shifting with difficulty level, can exist, we would suggest the reserving of the name *difficulty* factors for them. If factors arising from the application of linear equations to skewed data (although we can see no useful purpose in doing this) are going to be obtained, we suggest that they be called *constant error* factors to distinguish them from *content* factors.

Since tests chosen for factor analysis problems according to some extraneous psychological criterion are apt to vary in difficulty, as shown by their frequently skewed distributions, it is recommended that factor analysis studies should always be carried out on dichotomized test scores with intercorrelations computed by the tetrachoric formula. An alternative is to use only tests that yield *continuous normal* distributions, in which case product-moment r's would apply, but this might be neither possible nor advantageous.

References

1. Fergeson, G. A. The factorial interpretation of test difficulty. *Psychometrika*, 1941, **6**, 323-329.
2. Guilford, J. P. The difficulty of a test and its factor composition. *Psychometrika*, 1941, **6**, 67-77.
3. Richardson, M. W. The relation between the difficulty and differential validity of a test. *Psychometrika*, 1936, **1**, 33-49.
4. Walker, D. A. Answer pattern and score scatter in tests and examinations. *Brit. J. Psychol.*, 1931, **22**, 73-86; 1936, **26**, 301-308; 1940, **30**, 248-260.

PSYCHOMETRIKA—VOL. 9, NO. 4
DECEMBER, 1944

THE INTERPRETATION OF A TEST VALIDITY COEFFICIENT IN TERMS OF INCREASED EFFICIENCY OF A SELECTED GROUP OF PERSONNEL

LT. COL. MARION W. RICHARDSON
PERSONNEL RESEARCH, CLASSIFICATION AND REPLACEMENT BRANCH
ADJUTANT GENERAL'S OFFICE, WASHINGTON, D. C.

The predictive efficiency of a test used to select personnel is defined in terms of total effectiveness of the group thus selected, as compared with chance selection. The formula developed requires the use of an estimate of the ratio of average effectiveness of men selected to the average effectiveness of men not selected by the test. The predictive efficiency of the test varies directly with the magnitude of this ratio and also directly with the percentage rejected.

The validity coefficient, which is ordinarily the coefficient of correlation between a test (or test battery) and a criterion of effectiveness on the job, does not lend itself to an interpretation that is scientifically sound and at the same time meaningful to the average "practical" executive or personnel officer. The measures of predictive efficiency that are used by statisticians ordinarily involve an account of the relationship between the observed variance of a criterion group and the variance of measures of predicted effectiveness on the job. Such measures can only with great difficulty be explained to persons untrained in the rudiments of statistical theory. It is the purpose of this paper to develop and describe a measure of predictive efficiency that can be interpreted by practically anyone.

In addition to ease of interpretation, the measure must involve no special assumptions (such as normality) with respect to the shape of the bivariate test-criterion distribution, and must take account of the varying proportions of men regarded as successful (or satisfactory) in any criterion sample group.

Let us consider that the criterion sample group can be divided into two groups, "Satisfactory" and "Unsatisfactory," and that P_2 per cent of the group is classified as satisfactory. The percentage of unsatisfactory workers is therefore $1 - P_2$. Let us, then, designate the percentage of persons passing the test (or test battery) at an acceptable level as P_1. Let d equal the percentage of persons passing the test and regarded as satisfactory by the criterion. These definitions give us the following four-fold frequency table:

		Test		
		Fail	Pass	
Criterion	Satisfactory	$P_2 - d$	d	P_2
	Unsatisfactory	$1 - P_1 - P_2 + d$	$P_1 - d$	$1 - P_2$
		$1 - P_1$	P_1	1

Now let k equal the ratio of the average effectiveness of the P_2 men regarded as satisfactory to the average effectiveness of those men regarded as unsatisfactory by the criterion. The postulation of the factor k is not dependent on the possibility of an accurate estimate of its value, but, of course, the usefulness of the final formula is affected by the accuracy of such estimate.

The average effectiveness of men in the entire group is then $kP_2 + 1 - P_2$. The average effectiveness of men selected by the test is $\dfrac{kd + P_1 - d}{P_1}$, where $k \geq 1$. The percentage increase in effectiveness is given by

$$E = \frac{\dfrac{kd + P_1 - d}{P_1} - (kP_2 + 1 - P_2)}{kP_2 + 1 - P_2}, \tag{1}$$

wherefrom we can write:

$$E = \frac{(k-1)(d - P_1P_2)}{P_1(kP_2 - P_2 + 1)}. \tag{2}$$

The four-fold Pearsonian coefficient of correlation is not affected by the scalar k introduced, and is written as usual:

$$r = \frac{d(1 - P_1 - P_2 + d) - (P_1 - d)(P_2 - d)}{\sqrt{P_1(1 - P_1)P_2(1 - P_2)}}, \tag{3}$$

which may be simplified to

$$r = \frac{d - P_1P_2}{\sqrt{P_1P_2(1 - P_1)(1 - P_2)}}. \tag{4}$$

Substituting $r\sqrt{P_1P_2(1 - P_1)(1 - P_2)}$ for $d - P_1P_2$ in equation (2), we have

$$E = \frac{r(k-1)\sqrt{P_1P_2(1-P_1)(1-P_2)}}{P_1(kP_2 - P_2 + 1)}. \tag{5}$$

If the special case in which $P_1 = P_2 = P$ is admitted as not detracting seriously from the generality of the formula, equation (5) becomes simply

$$E = \frac{r(k-1)(1-P)}{P(k-1)+1}. \tag{6}$$

The formula gives values of 0% to 100% in increased percentage effectiveness over chance selection for values of $k = 1$ to $k = \infty$, the values of E increasing in a negatively accelerated manner. The values of E decrease as P increases. The value P has been called the selection ratio. Taylor and Russell* have computed, on the basis of an assumed normal bivariate distribution, the proportions of men who will be satisfactory among those selected for given values of the selection ratio and the validity coefficient. Their tables show the formal fact that, for given validity coefficients, the selection is increasingly effective as the selection ratio is lowered. The value of E may be greater than 100% for sufficiently small values of P.

The use of formula (6) can be illustrated by a simple example, not atypical of prediction in a personnel situation. Test X correlates .52 with a criterion of job effectiveness in a labor market such that the upper 25 per cent of applicants can be selected. The criterion measures indicate that the best 25 per cent of the criterion group is $3\frac{1}{2}$ times as effective as the lower 75 per cent of the group. The percentage of increased effectiveness attributable to the use of the test is given by

$$E = \frac{.52(3.5-1)(.75)}{.25(3.5-1)+1} = 60\%.$$

The result is readily interpreted as the increase in efficiency due to the use of the test to select the men needed as compared with selecting the men at random.

The formula as developed may be applied to interviews, scored personal history blanks, or other selection devices in the same manner as tests. The principal difficulty in the use of the formula lies in the availability of k as a datum. Where the work performed by the personnel is of a simple clerical or mechanical nature and the pieces

* Taylor, H. C. and Russell, J. T. The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. *J. appl. Psychol.*, 13, 1939, 565-578.

can be counted, an estimate of k is simple. In certain school situations, for example, radio code training schools, an estimate of k may be made in terms of the ratio of time required to reach a given code speed.

For many occupations where products cannot be counted, the estimate will be difficult. A consensus of supervisors might be resorted to, with the understanding that the estimate would have the inaccuracy characteristic of all rating methods. Moreover, supervisors tend to underestimate individual differences in the efficiency of their employees.

PSYCHOMETRIKA—VOL. 9, NO. 4
DECEMBER, 1944

MATHEMATICAL ANALYSIS IN PSYCHOLOGY OF EDUCATION: COMPUTATION OF STIMULATION, RAPPORT, AND INSTRUCTOR'S DRIVING POWER

MICHAEL A. SADOWSKY*
ILLINOIS INSTITUTE OF TECHNOLOGY

Mathematical expressions are derived for such concepts as stimulation of student by instructor, student-instructor rapport, and driving power of instructor, in terms of the student's and the instructor's foci of attention, their strength of concentration, and the intensity of the presentation and of the reception of details of subject matter. Under the assumption of normal distribution, the mathematical methods of combination and integration yield conclusions on summary integral effects of interrelations within the educational team. The psychological interpretation of the mathematical results thus obtained conforms with common sense. The main emphasis of the article is the exposition of how the mathematical method of combination and integration can be used to estimate the resultant effect of various independent combined simple factors acting independently within the individuals forming the educational team. No claim is made as to the absolute truthfulness and reliability of the psychological postulates used at the beginning stage of the mathematical analysis.

1. *Introduction*

The author, an applied mathematician engaged in both teaching and research, recently became involved in work in the psychology of education. He has occasionally overheard students say "instructor A is better than instructor B," occasionally an instructor say "student C is better than student D." The comparative "is better," to make sense, must be equivalent to inequalities $A > B$, $C > D$ which implies the existence of a linear scale of distribution for A, B, C, D. Being in complete agreement with the general terminology in psychology, the author has been longing for a numerical treatment of the issues

* In presenting this paper the primary intention of the author has been to apply certain mathematical techniques to the field of psychology. The method presented is illustrative and the mathematical techniques are the main emphasis of the article. The absolute truthfulness of the psychological postulates used and of their implications will doubtless be questioned by readers of higher competence in the field of psychology. If psychological postulates of less uncertain and less questionable character should be substituted by experts in the field to lay a basis for the mathematical analysis, the psychological content and the psychological implications derived should doubtless be on a higher level of credibility.

The author wishes to express his indebtedness to Mr. Willard Skolnik of the Illinois Institute of Technology for his valuable suggestions regarding the form of presentation of the material involved.

involved in psychology of education and now undertakes an approach from the numerical side. In showing by means of the mathematical mechanism developed that mentally limited students ("single-track minds," mathematically: one-dimensional minds) may not profit from instruction offered in excess of 41.4% on a certain scale (section 9, equation 15), while mentally many-sided students may profit to the extent of several hundred per cent (section 10, table 18, equation 17), the author found his mathematical method in agreement with his personal experience (and, probably, everybody else's experience) and yields to an urge to present the method of mathematical analysis in psychology to the attention of psychologists and educators.

2. Notations and Terminology

x .... a real variable used to represent the entire contents of the subject matter

x_1, x_2, x_3 ... particular separate details (facts, statements) within the scope of the subject

x_1 .... focus of concentration of instructor's attention

x_2 .... focus of concentration of student's attention

if $x_1 = x_2$ state of complete rapport established

if $x_1 \neq x_2$ rapport is incomplete

s_1 .... instructor's strength of concentration

s_2 .... student's strength of concentration

$f_1(x)$ .... intensity of presentation of the detail x by an instructor whose attention is focused at x_1

$f_2(x)$ .... intensity of reception of the detail x by a student whose attention is focused at x_2

S .... total stimulation originated within the student by the instructor (or stimulation received by the student from the instructor)

R .... maximum value of total stimulation originated within the student by the instructor in state of complete rapport

r .... relative rapport ($r = 1$ for complete rapport)

D .... driving power of the instructor as felt by the student (drive)

n .... number of independent ("orthogonal") dimensional components within the entire subject

3. Setting the Stage

Let us consider a class in a well defined subject of instruction led by an instructor in an intellectual way demanding uninterrupted

attention and unceasing mental cooperation of the class. Let us single out one student for further reference. Assuming cooperation on both parts, we may describe the performance done in the teaching-learning process as follows:

4. Individual Performance of the Instructor

The instructor acts as a source of visible and audible stimulation related to a definite subject whose contents we will represent by a real variable x. Different single facts, details, statements, etc., within the subject x will correspond to different particular values $x_1, x_2, x_3, \cdots$ of the variable x. At a given moment, the instructor's activity will be centered about some detail, say x_1, within the scope of the total subject x. He will concentrate all of his consciousness, attention, interest, mental preoccupation about the "focus of concentration" x_1. Let us assume that the law of normal distribution holds for the intensity of presentation $f_1(x)$ of any other related detail x appearing by association in the presentation by the instructor whose attention is focused at x_1. The law of normal distribution gives

$$f_1(x) = \sqrt{2}\, s_1\, e^{-2\pi s_1^2 (x-x_1)^2}. \tag{1}$$

s_1 is a personal constant determined by the instructor's individuality. It expresses the strength of concentration of the instructor on the focus x_1 within the scope of the entire subject x. For explanation, see section 8. Mathematically speaking, the range of possible values of the strength of concentration is unlimited:

$$0 \leq s < \infty . \tag{2}$$

Assuming competence on the part of the instructor, we expect a large value of his strength of concentration s_1. This gives the "sharp" type of normal distribution with a high salient at the focus of concentration at the object of discussion x_1 and insignificant dissipation of interest toward distant details.

The total amount of instructor's attention to the subject x, while discussing the detail x_1, is 100% according to

$$\int_{-\infty}^{\infty} f_1(x)\, dx = 1 . \tag{3}$$

Please note that this is instructor's attention to the subject—it is not his attention to the reception of the subject by the student.

5. Individual Performance of the Student

The student, listening as an independent individual, exercises

his receptive attention. His mind may be focused about a point of interest x_2, the same as or other than the instructor's focus x_1. The distribution of the student's intensity of receptive attention as expressed by a normal law with a personal constant s_2 is

$$f_2(x) = \sqrt{2}\, s_2\, e^{-2\pi s_2^2 (x-x_2)^2}. \tag{4}$$

s_2 will be referred to as the student's strength of concentration on the focus x_2 within the scope of the entire subject x. The total amount of the student's attention to the subject x, while having his attention focused at x_2, is 100% according to

$$\int_{-\infty}^{\infty} f_2(x)\, dx = 1\,. \tag{5}$$

Please note that this is the student's attention to the subject—it is not his attention to the presentation of the subject by the instructor.

The "sharpness" of the normal curve depends on the value of s_2. A small value of s_2 would give a low-level curve with no recognizable concentration of attention at the focus (attempts to concentrate thought on the focus fail, no prominent intensity of thought is reached in the region of the focus and innumerable distant associations (dissociations) are blurring the mental picture). In case of a small s_2, the total 100% of attention is indiscriminately distributed over the range of the subject x. The opposite type arising with a large value of s_2 was discussed in section 4.

6. *The Teaching-Learning Team*

In sections 4 and 5 we have dealt individually and separately with the instructor and the student, crediting each one with 100% of attention to the total subject matter. Both have been introduced on equal terms as two indispensable partners in a team. We now progress to an analysis of the team work. The basic problem here is the computation of the stimulation as received by the student from the instructor.

Let the instructor whose focus of attention is at x_1 discuss at a given moment a detail covering the range from x to $x + dx$ on the subject axis of x. Since the instructor's intensity of presentation of the subject at the place x is $f_1(x)$, the stimulation given out by the instructor will amount to $f_1(x)\,dx$. Since the student's intensity of receptive attention at the place x is $f_2(x)$, the stimulation actually received by the student is $f_1(x) f_2(x)\,dx$. The total stimulation S received by the student is obtained by integration over the entire subject as

$$S = \int_{-\infty}^{\infty} f_1(x) f_2(x)\, dx$$
$$= \sqrt{2} \frac{s_1 s_2}{\sqrt{s_1^2 + s_2^2}} e^{-2\pi \frac{s_1^2 s_2^2}{s_1^2 + s_2^2} (x_2 - x_1)^2} . \tag{6}$$

7. *Rapport*

If the student and the instructor concentrate their attention at different foci ($x_1 \neq x_2$), we may say that the rapport between them is incomplete: in the student's opinion, the instructor does not approach the subject from the right point of view, while in the instructor's opinion the student does not react in the right way. If both foci of attention coincide, that is,

$$x_1 = x_2 \tag{7}$$

unity of interest is present and complete rapport has been established between student and instructor. Let us denote by R the value of stimulation S in a state of complete rapport. Putting $x_1 = x_2$ in equation (6), we obtain

$$R = \sqrt{2} \frac{s_1 s_2}{\sqrt{s_1^2 + s_2^2}} . \tag{8}$$

Equation (6) may now be re-written as

$$S = R \cdot e^{-\pi R^2 (x_2 - x_1)^2} . \tag{9}$$

We see that the value of stimulation in complete rapport, R, can be determined if the values s_1 and s_2 are known, and that maximum stimulation between two given individuals is reached in case complete rapport is established.

To measure the value of the incomplete rapport on a relative basis, let us use for this purpose the ratio of the stimulation S actually received by the student in case of incomplete rapport and stimulation R in case of complete rapport. Using r to denote this ratio and referring to it as relative rapport, we obtain

$$r = \frac{S}{R} = \frac{\text{stimulation actually received in incomplete rapport}}{\text{maximum stimulation possible in complete rapport}} . \tag{10}$$

From equations (9) and (10) we have as the final expression for the relative rapport

$$r = e^{-\pi R^2 (x_2 - x_1)^2} . \tag{11}$$

The maximum value of the relative rapport is unity (= 100%), which is reached in case complete rapport is established.

Note: From equation (11) follows $(x_2 - x_1)^2 = -\dfrac{\log r}{\pi R^2}$, which shows that the amount of disturbance of attention, $x_2 - x_1$, depressing an established complete rapport from the relative value of unity down to r, varies inversely with the absolute value R of the complete rapport. The difficulty of keeping up a complete rapport increases with its absolute value. This "trivial" (part of everybody's experience) statement has been obtained here as a consequence from equation (11).

8. *Individual Self-Stimulation*

Let us consider a situation in which both partners of the dual team have been separated and each one isolated to perform some mental work requiring deep thinking within the entire extent of the subject x, using no reference, obtaining no help from outside whatsoever. The individual positions of both partners are basically equivalent: each one is engaged in a two-fold process: (1) to produce by recall a conscious mental display of the subject x, and (2) to perceive that display as an observer and use it as a stimulus for further progressive reasoning. Such a process of contemplation and reasoning is comparable to an impersonation of "instructor" and "student" in a single individual in either case (personal union, "teaching oneself"). Assuming that the focus of concentration in recall is identical with the focus of concentration in reception (no split attention), we will find the person in a state of complete rapport with himself. We further assume that the personal constant s_1 applies to the instructor for both intensity of recall and intensity of perception, while s_2 is valid for the student both ways in the same sense. The value of stimulation the person can originate within himself is computable from the rapport equation (8) in which, if that person is the instructor, s_2 is to be replaced by s_1, giving

$$R = s_1 \tag{12}$$

as the value of self-stimulation of the instructor. If the person is the student, s_1 is to be replaced by s_2, giving

$$R = s_2 \tag{13}$$

as the value of self-stimulation of the student. Hence, in any case, the value of self-stimulation that a person may originate within himself by himself is equal to the strength of concentration constant of

that person. Considering the unbounded range of s (equation 2), we see that considerable personal differences may exist in the amount of self-stimulation possible.

9. *Driving Power of the Instructor*

In establishing rapport the instructor's objective is to maximize the value of stimulation received by the student according to equation (11). The instructor has to drive the student into acceptance of his, the instructor's, focus of concentration x_1. The acceptance by the student of x_1 as focus of concentration is psychologically engineered by a performance on the part of the instructor which is useless for any aim other than concentration on x_1 (large value of instructor's s_1 in presentation of the subject). If the student is still able to follow and to cooperate mentally, he could measure the driving power of the instructor, D, by the stimulation which he receives from the instructor relative to the self-stimulation which he could produce without the instructor. The student's stimulation in case of complete rapport is R (equation 8), his self-stimulation is s_2 (equation 13). Dividing R by s_2, we obtain for the value of the drive (driving power of the instructor)

$$D = \sqrt{2}\frac{s_1}{\sqrt{s_1^2 + s_2^2}} = \sqrt{\frac{2}{1 + \left(\frac{s_2}{s_1}\right)^2}} . \tag{14}$$

A value of D greater than unity is apt to produce in the student feelings of actually being helped by instruction (precision of thought, clarification of thought, security and efficiency of thinking, realization of aim, etc.) A value of D less than unity would produce the opposite feelings, such as "mental fog," "things messed up by instructor," "lost track," "ceasing to understand instructor at all," etc.

By equation (14) the drive increases as the ratio s_2/s_1 decreases. To serve a given student best (s_2 is given), the instructor engaged should have $s_1 = \infty$ (superman). In this case, the drive attains its maximum value of

$$D_{\max} = \sqrt{2} = 1.414 , \tag{15}$$

which shows but a finite increase of 41.4% of the stimulation of the student—in spite of the presence of a super-instructor as the second member of the team. The question of dimensionalities as discussed in the section following shows that those 41.4% may or may not be the ultimate limit, depending on the number, n, of dimensions within the subject of instruction.

10. Effect of Dimensionality

The results obtained so far were based on the assumption that the contents of the subject taught could be continuously displayed along the axis of a single real variable x (section 4). If such is not the case, a distribution of the subject of two, three, or more dimensions ought to be used. Multiple integration would replace single integration, and the resultant driving power D of the instructor within the entire multi-dimensional subject will be the product of his partial driving powers, $D_1, D_2, D_3, \cdots$ within each component of the subject:

$$D = D_1 \cdot D_2 \cdots D_n . \tag{16}$$

Each partial drive $D_1, D_2, D_3, \cdots$ would be computable from equation (14), and the value of any partial drive would never exceed $\sqrt{2}$. Consequently, in case of n dimensions, the maximum value of the drive (super-instructor superior in all dimensions) is given by

$$D_{\max} = 2^{n/2} \tag{17}$$

which can be tabulated as follows:

$n =$	1	2	3	4	5
$D_{\max} =$	1.41	2.00	2.83	4.00	5.66
Maximum increase of stimulation	41%	100%	183%	300%	466%

. (18)

11. Conclusion

In section 1 (Introduction) the following statement was made: "Mentally limited students ("single-track minds," mathematically: one-dimensional minds) may not profit from instruction offered in excess of 41.4% on a certain scale, while mentally many-sided students may profit to the extent of several hundred per cent." This statement is now explained by linking "profit from instruction" with the difference $D - 1$ (last row in table 18) and many-sidedness of a student with the number n of active dimensions which he is able to actively engage as independent avenues of approach in his penetration into the subject of instruction.

PSYCHOMETRIKA—VOL. 9, NO. 4
DECEMBER, 1944

A SIMPLE METHOD OF FACTOR ANALYSIS

KARL J. HOLZINGER
THE UNIVERSITY OF CHICAGO

A simple method for extracting correlated factors simultaneously is described. The method is based on the idea that the centroid pattern coefficients for the sections of unit rank of the complete matrix may be interpreted as structure values for the entire matrix. Only the routine centroid average process is required.

The simple method here presented is applicable to the factoring of a correlation matrix in case the latter can be sectioned into portions of approximate rank unity. Such sectioning can often be accomplished by inspection of the whole matrix, by the use of B-coefficients,* and by the nature of the variables involved.

After the matrix is sectioned, centroid coefficients for the variables in each section are computed. These coefficients may be based upon communalities or other values in the diagonals. For any *one section* the first centroid coefficients a_{js} may be interpreted as pattern values† or as *structure* values, since they show correlation between a variable z_j and the centroid C_s of a particular section. The first centroids of the various sections will be correlated, however, and the coefficients a_{js} extended to all variables may then be interpreted as structure values of s_{js} for the entire correlation matrix. It is this simple idea that is the basis of the present method.

The adequacy of the solution may be tested in each section, regarding the a_{js} of that section as pattern values which will yield pattern correlations for a given section. A complete pattern may also be found if desired in order to test for overlapping of factors in the remainder of the correlation matrix.

The reproduced correlations in a section are

$$r'_{jk} = a_{js} a_{ks} , \tag{1}$$

where the factor pattern of a section has the form,

* Karl J. Holzinger and Harry H. Harman, Factor analysis, p. 24. Chicago: University of Chicago Press, 1941.

† *Ibid.*, p. 16. A "structure" S_{js} is a matrix of correlations between tests and factors.

$$\left.\begin{array}{l} z_1 = a_{1s}C_s \\ z_2 = a_{2s}C_s \\ \cdot\ \cdot\ \cdot\ \cdot \\ \\ z_p = a_{ps}C_s \end{array}\right\} \begin{array}{l} j = 1, 2, 3, \cdots, p. \\ p = \text{number of variables in a section.} \\ s = 1, 2, 3, m. \\ m = \text{number of centroids.} \\ \\ \text{Specific factors not shown.} \end{array} \tag{2}$$

A simple discussion of the centroid method as applied here will next be given. Assume that the variables z_1, z_2, and z_3 are in a given section and that z_u is not. The theoretical pattern would then have the form,

$$\begin{array}{l} z_1 = a_1C \\ z_2 = a_2C \\ z_3 = a_3C \\ \hline z_u = a_uC \end{array} \tag{3}$$

wherein the second subscript has been dropped for simplicity.

The correlation matrix for pattern (3) may be written

$$\begin{array}{rccc|c|c} R = & a_1a_1 & a_1a_2 & a_1a_3 & & a_1a_u \\ & a_2a_1 & a_2a_2 & a_2a_3 & & a_2a_u \\ & a_3a_1 & a_3a_2 & a_3a_3 & & a_3a_u \\ \hline \text{Sum} & T_1 & T_2 & T_3 & T = T_1 + T_2 + T_3 & T_u \end{array} .$$

The centroid coefficient for z_1 may then be written in the form,

$$a_1 = \frac{T_1}{\sqrt{T}} = \frac{a_1(a_1 + a_2 + a_3)}{\sqrt{a_1(a_1 + a_2 + a_3) + a_2(a_1 + a_2 + a_3) + a_3(a_1 + a_2 + a_3)}}$$

$$= \frac{a_1(a_1 + a_2 + a_3)}{\sqrt{(a_1 + a_2 + a_3)^2}} = a_1 = \frac{r_{11} + r_{12} + r_{13}}{\sqrt{\text{Sum of } 9r\text{'s}}}. \tag{4}$$

The coefficient for z_u is taken as

$$a_u = \frac{T_u}{\sqrt{T}}, \tag{5}$$

employing the T for z_1, z_2, z_3 as before. It is therefore apparent that within the group a_1, a_2, a_3 are both pattern and structure values, whereas the values a_u are structure values for the whole matrix. For this reason the values from formulas (4) and (5) are denoted as s_{js} in Table 1. The structure is required in case estimates of the factors are desired. The calculations may be made by the Doolittle* or similar method.

* *Ibid.*, p. 390.

A complete outline for the calculation is given on the work sheet, using the data of Table 7.1 for eight physical variables* with bifactor communalities from Table 8.4.† The matrix has been sectioned (1,2,3,4), (5,6,7,8), as shown on the work sheet.

After the values s_{js} have been computed, they may be arranged as in the following table:

TABLE 1
Common Structure S_{js}

Variable	s_{j1}	s_{j2}
1	.919	.484
2	.942	.434
3	.907	.399
4	.893	.454
5	.455	.932
6	.375	.813
7	.312	.739
8	.412	.724

These values agree with those of Table 11.2 within rounding error of .001.‡

The fit of the correlations for the sections yielding the factors may be obtained from the reproduced correlations at the bottom of the work sheet. The residuals in general are small.

In case the complete pattern is required, it is necessary to obtain the correlation between factors. This correlation may be found from the intercorrelations of the variables in the inter-group sections. In the present example the intercorrelations among variables in the groups (1,2,3,4) and (5,6,7,8) would be employed. The average of these sixteen values, using the subtotals from the work sheet, is

$$\frac{1.554 + 1.394 + 1.281 + 1.457}{16} = .3554.$$

This last average must be corrected, however, in order to obtain the values for the correlations in the common-factor space.‖

The values $s_{11} = .919$, $s_{21} = .942$, $s_{31} = .907$, and $s_{41} = .893$ may be interpreted as the lengths of the vectors z_1, z_2, z_3, and z_4 projected on the $0C_1$ axis. Their average length is

* *Ibid.*, p. 169.
† *Ibid.*, p. 192.
‡ *Ibid.*, p. 245.
‖ *Ibid.*, p. 61. See formula 3.50, which is here applied to averages of variables.

$$\frac{.919+.942+.907+.893}{4}=.9152.$$

Similarly the average length of the vectors z_5, z_6, z_7, and z_8 projected on the $0C_2$ axis is

$$\frac{.932+.813+.739+.724}{4}=.8020.$$

The correlation between factors C_1 and C_2 in the common-factor space is then

$$r_{C_1C_2}=\frac{.3554}{.9152\times .8020}=.4842.$$

From the structure S_{js} and the correlation between factors ϕ_{ss} the pattern may be found by the method of Appendix G.3.*

TABLE 2
Common Pattern A_{js}

Variable	a_{j1}	a_{j2}
1	.894	.051
2	.956	—.029
3	.932	—.052
4	.879	.029
5	.005	.930
6	—.024	.824
7	—.060	.768
8	.080	.685

It is now possible to check the fit of the whole observed correlation matrix by obtaining the complete matrix of reproduced correlations. Denoting this matrix as R^+, then

$$R^+_{jj}=A_{js}\phi_{ss}A'_{js}\,, \tag{6}†$$

but since

$$S_{js}=A_{js}\phi_{ss}\,, \tag{7}‡$$

then

$$S'_{js}=\phi_{ss}A'_{js} \tag{8}$$

and

$$R^+=A_{js}S'_{js}=S_{js}A'_{js}\,. \tag{9}$$

* *Ibid.*, p. 386.
† *Ibid.*, p. 19.
‡ *Ibid.*, p. 327. (Here, T instead of S denotes structure.)

This last equation is very convenient for obtaining the reproduced correlations inasmuch as the correlation ϕ_{ss} is not explicitly required. From Tables 1 and 2 the product $S_{js}A'_{js}$ is found to be

$$\left\| \begin{matrix} .846 & .865 & .831 & .822 & .455 & .377 & .317 & .405 \\ .864 & .888 & .855 & .841 & .408 & .335 & .277 & .373 \\ .831 & .856 & .825 & .809 & .376 & .307 & .252 & .346 \\ .821 & .841 & .809 & .798 & .427 & .353 & .295 & .382 \\ .454 & .408 & .376 & .427 & .869 & .757 & .688 & .675 \\ .377 & .335 & .307 & .353 & .758 & .661 & .602 & .587 \\ .317 & .277 & .252 & .296 & .689 & .601 & .549 & .531 \\ .405 & .373 & .346 & .383 & .675 & .587 & .531 & .529 \end{matrix} \right\|$$

Upon comparing the original correlation matrix given on the work sheet with the above reproduced correlations, it is apparent that all the residuals are negligible. The fit of the pattern is therefore an excellent one.

It will be observed that the present method obviates the usual procedure of first obtaining an orthogonal centroid solution for the entire correlation matrix, and then rotating to oblique axes satisfying what Thurstone would call a special case of "simple structure." Such "simple structure" is tested here analytically by checking the rank of the submatrixes as shown at the bottom of the work sheet, and by testing the goodness of fit of the entire correlation matrix from equation (9). If large residuals occur in either case, the variables may be rearranged and the correlation matrix resectioned to secure a better fit. There is no guarantee, of course, that any correlation matrix can be factored in the above manner, but if either a bi-factor pattern or this type of "simple structure" exists, then the above method is applicable and is much more simple and direct than those now in use.

WORK SHEET FOR SIMPLE METHOD

	1	2	3	4		5	6	7	8
1	.854	.846	.805	.859		.473	.398	.301	.382
2	.846	.897	.881	.826		.376	.326	.277	.415
3	.805	.881	.833	.801		.380	.319	.237	.345
4	.859	.826	.801	.783		.436	.329	.327	.365
Sum	3.364	3.450	3.320	3.269	$T = 13.403$	1.665	1.372	1.142	1.507
a_{1j}	.919	.942	.907	.893	$\sqrt{T} = 3.661$	.455	.375	.312	.412
					$\frac{1}{\sqrt{T}} = .2731$				

5	.473	.376	.380	.436		.870	.762	.730	.629
6	.398	.326	.319	.329		.762	.687	.583	.577
7	.301	.277	.237	.327		.730	.583	.521	.539
8	.382	.415	.345	.365		.629	.577	.539	.579
Sum	1.554	1.394	1.281	1.457		2.991	2.609	2.373	2.324
	.484	.434	.399	.454	a_{2j}	.932	.813	.739	.724

$$T = 10.297 \qquad \sqrt{T} = 3.209 \qquad \frac{1}{\sqrt{T}} = .3116$$

	a_{j1}	$\times$	a'_{j1}				$=$		R'			
1	.919	$\times$	.919	.942	.907	.893	$=$	1	.845	.866	.834	.821
2	.942							2	.866	.887	.854	.841
3	.907							3	.834	.854	.823	.810
4	.893							4	.821	.841	.810	.797

	a_{j2}	$\times$	a'_{j2}				$=$		R'			
5	.932	$\times$	.932	.813	.739	.724	$=$	5	.869	.758	.689	.675
6	.813							6	.758	.661	.601	.589
7	.739							7	.689	.601	.546	.535
8	.724							8	.675	.589	.535	.524

PSYCHOMETRIKA—VOL. 9, NO. 4
DECEMBER, 1944

MAXIMAL WEIGHTING OF QUALITATIVE DATA

ROBERT J. WHERRY
UNIVERSITY OF NORTH CAROLINA

A method whereby biographical or other questionnaire data of a purely qualitative nature may be used to predict success or failure on an independent criterion is presented. The method is not new but the present least-squares derivation and the transformation equation for punched card coding were not available in the literature. The proper weights are found to be proportional to the per cent of passers in the various categories. The method is suggested as a suitable substitute for non-linear approaches in connection with purely quantitative data as well. The implications of reweighting in connection with multiple regression is discussed. The lavish use of degrees of freedom makes cross-validation extremely desirable.

The procedure to be described is one that the present writer acquired at the Ohio State psychological laboratory several years ago and had accepted "on faith." When the usage was questioned lately, he could find no derivation or even any mention of such a method in the literature. The present least squares derivation may make the method acceptable and add a useful technique to the repertoire of others not hitherto familiar with this procedure.

The problem arises when biographical or other questionnaire data of a purely qualitative nature are to be used to predict success or failure on an independent criterion. Mere presence or absence of relationship can of course be established by the Chi-squared (contingency) approach or by analysis of variance, but if actual prediction is desired and the number of categories is greater than two, numerical weights must be assigned to each such category.

As an example let us consider the following general scatter diagram:

	Criterion			
Category	fail	pass	total	
a	fr_{qa}	fr_{pa}	fr_{ta}	
b	fr_{qb}	fr_{pb}	fr_{tb}	
c	fr_{qc}	fr_{pc}	fr_{tc}	(1)
.	. . .	. . .	. . .	
n	fr_{qn}	fr_{pn}	fr_{tn}	
$\Sigma =$	Nq	Np	N	

We want to assign quantitative weights to the qualitative categories $a, b, c, \cdots, n$ such that the resulting bi-serial or point-biserial correlation will be a maximum.

The formula for the correlation coefficient is

$$r = \frac{M_p - M_q}{\sigma_t} \cdot K, \tag{2}$$

where M_p = mean of the passers on the multi-categoried variable,
M_q = mean of the failers on the multi-categoried variable,
σ_t = standard deviation of the multi-categoried variable,
and $K = pq/z$ (if biserial) or $\sqrt{pq}$ (if point-biserial) on the criterion variable.

Thus, in terms of the entries in set (1), we have

$$r = \frac{\dfrac{\sum_a^n X_u fr_{pu}}{Np} - \dfrac{\sum_a^n X_u fr_{qu}}{Nq}}{\sqrt{\dfrac{\sum_a^n x_u^2 fr_{tu}}{N}}} \cdot K \tag{3}$$

or, since $X_u = x_u + M_u$, we can reduce all scores to a deviation basis, getting

$$r = \frac{\dfrac{\sum_a^n (x_u + M_u) fr_{pu}}{Np} - \dfrac{\sum_a^n (x_u + M_u) fr_{qu}}{Nq}}{\sqrt{\dfrac{\sum_a^n x_u^2 fr_{tu}}{N}}} \cdot K. \tag{4}$$

Of course the x_u's are the unknown values to be determined so as to make the r a maximum. First we take the logarithm of both sides of equation (4), obtaining

$$\begin{aligned} \log r &= \log[\textstyle\sum_a^n (x_u + M_u) fr_{pu}/Np - \sum_a^n (x_u + M_u) fr_{qu}/Nq] \\ &\quad - 1/2 \log(\textstyle\sum_a^n x_u^2 fr_{tu}/N) + \log K \\ &= \log A - 1/2 \log B + \log K . \end{aligned} \tag{5}$$

To maximize $\log r$, and thus r, we take the partial derivatives of $\log r$ with respect to the unknowns $x_a, x_b, \cdots, x_n$, set these derivatives equal to zero, and solve for the values of x_u. Thus we have

$$\begin{aligned} \delta \log r/\delta x_u &= \delta \log A/\delta x_u - 1/2\, \delta \log B/\delta x_u + \delta \log K/\delta_u \\ &= (fr_{pu}/Np - fr_{qu}/Nq)/A - x_u fr_{tu}/B + 0 = 0 . \end{aligned} \tag{6}$$

We have n such equations with the subscript u taking all values from a to n, and the general solution is

$$\begin{aligned}
x_u &= \frac{fr_{pu}/Np - fr_{qu}/Nq}{fr_{tu}} \cdot \frac{BN}{A} \\
&= \frac{q\,fr_{pu} - p\,fr_{qu}}{fr_{tu}} \cdot \frac{B}{Apq} \\
&= (q\,\%_{pu} - p\,\%_{qu}) \cdot B/100Apq \\
&= [(p+q)\,\%_{pu} - 100p] \cdot B/100Apq \\
&= \%_{pu} \cdot B/100Apq - B/Aq \\
&= k_i\,\%_{pu} - k_{ii}\,,
\end{aligned} \tag{7}$$

and since k_i and k_{ii} are constants for all values of u (for each category) we may write simply

$$\begin{aligned}
x_u(\text{proportional}) &= \%_{pu} \\
&= \text{Percentage passing in the category.}
\end{aligned} \tag{8}$$

Since these weights are already relative they can be transformed with only infinitesimal loss of accuracy into a series ranging from 0 to 9 for IBM punched card operations by the following transformation equation:

$$X_{u\ (\text{Code})}' = 9\left[\frac{\%_{pu} - \%_{p(\text{lowest})}}{\%_{p(\text{highest})} - \%_{p(\text{lowest})}}\right]. \tag{9}$$

As an example consider the following fictitious situation*

Category	Major in College	Navigator fail	Navigator pass	$x_u = \%_{pu}$	$X_u' =$ IBM code
a	Engineering	2	8	80	9
b	Chemistry	10	20	67	7
c	Mathematics	12	36	75	8
d	English	4	1	20	0
e	Commerce	10	15	60	6
f	Philosophy	8	4	33	2
g	History	6	4	40	3

Where the coded primed scores were arrived at by the use of equation (9) as indicated below:

* It is obvious that the frequencies in the example are too few in most categories to establish reliable weights, but the principle remains the same. In practice the usual formulas for reliability of percentages and the usual restrictions concerning cell frequencies in any contingency problems should be taken into account.

$$X_a' = 9(80\text{-}20)/(80\text{-}20) = 9.00 = 9$$
$$X_b' = 9(67\text{-}20)/(80\text{-}20) = 7.05 = 7$$
$$\cdots\cdots\cdots\cdots$$
and
$$X_n' = 9(40\text{-}20)/(80\text{-}20) = 3.00 = 3\,.$$

Criteron correlations may then be computed either as biserials or as point-biserials (straight Pearsonian with 0 and 1 for the fail-pass categories of the criterion) depending upon the proper assumption as to criterion distribution.

The intercorrelations between several such variables will indicate the degree of overlap (extent to which they correlate with the criterion by picking the same men in the high and low groups correctly) and will permit the calculation of multiple biserial or multiple point-biserial regression weights and correlation coefficients. Multiplying the primed weights by the regression weights to form a composite score will not affect them since they are already relative, but will merely weight the various tests (questions) so as to produce as little overlap as possible.

The writer has even found this technique useful in connection with the rescoring of quantified data in those cases where the relationship with the criterion was non-linear. If each class interval of the quantitative score series is given a primed score by the above technique the non-linearity vanishes and makes the usual correlation techniques adequate.

Since the method uses degrees of freedom rather lavishly, cross-validation is always extremely desirable.

PSYCHOMETRIKA—VOL. 9, NO. 4
DECEMBER, 1944

"PARALLEL PROPORTIONAL PROFILES" AND OTHER PRINCIPLES FOR DETERMINING THE CHOICE OF FACTORS BY ROTATION

RAYMOND B. CATTELL
DUKE UNIVERSITY

The choosing of a set of factors likely to correspond to the real psychological unitary traits in a situation usually reduces to finding a satisfactory rotation in a Thurstone centroid analysis. Seven principles, three of which are new, are described whereby rotation may be determined and/or judged. It is argued that the most fundamental is the principle of "parallel proportional profiles" or "simultaneous simple structure." A mathematical proof of the uniqueness of determination by this means is attempted and equations are suggested for discovering the unique position.

Source Traits or Mathematical Artifacts?

If factor analysis is used merely as a tool to obtain mathematical factors, a relatively small number of which will act as efficient predictors with respect to a relatively large number of individual variables, the problems discussed in this article do not arise. Any one set of mathematical "artifacts" is practically as good as another for prediction from any one test battery.

Psychological research, as such, cannot, however, be content with this limited goal. It strives toward psychologically meaningful functional unities.* In an earlier article (3) the present writer has sifted the uses of the term "trait" and has pointed out that there are only two legitimate senses in which a trait unity can be said to exist, (a) as a "surface trait" or correlation *cluster,* and (b) as a "source trait" or *factor.* Factor analysis then becomes considered by the psychologist as *the* device for discovering *source traits.* Since real source traits appear as factors, whereas not all factors represent real source traits, the problem next presents itself: "How can one decide which one, among many possible sets of factors, alone corresponds to the real functional unities in the psychological situation?" The senses in which the unitariness of a source trait can be considered more 'real' than that of a factor *per se* have been treated fully elsewhere (7).

* Reyburn and Taylor (13) and others have sometimes spoken of an intermediate degree of reality. "If common factors are not causal they must at least be objective . . . [which requires] a certain form and degree of invariance" [namely, of factor loading of a test from battery to battery]. We should consider such factors to be in a transitory limbo, destined soon to emerge to one status or the other.

Approaches to True Functional Unities

The discovery and confirmation of factors corresponding to real unities may be attempted, broadly, in two ways, as follows:

(1) The psychologist may set up, on psychological grounds, a precise hypothesis as to the number and nature of the source traits in the situation. If he can find among the many possible mathematical solutions of his factorizations one which exactly fits the hypothetical system, he may be said to have proved his hypothesis, as far as the coincidence of a set of facts from a hypothesis with observed facts ever proves a hypothesis.

(2) The psychologist may start out with no independent hypothesis. Instead he may start with general principles in the mathematical analysis which will lead to only one, uniquely-determined solution. This solution can also be used to confirm or destroy a hypothesis, but essentially it requires no hypothesis formation: it simply produces an observation, a discovery, the validity of which rests on the validity of the principles.

Because of the great difficulties so far experienced in getting generally acceptable and satisfactory principles for (2), many psychologists have announced themselves hopeless of ever achieving, through the agency of factor analysis, anything more than convenient mathematical artifacts. Others, with more moderate views, have fallen back wholly on (1). The latter, however, is a frail if not broken reed in most situations, for there may be so many mathematically possible factorizations, and our checks on sampling errors are so poor, that almost any hypothesis can be "satisfied." Occasionally the nature of the given battery and resultant correlation matrix limits the possible factors, e.g., to a general and specific factor, as in a matrix which happens to satisfy Spearman's criterion (14), or to positive factors only. In other cases one can also analyze into a general factor and non-overlapping group factors, as by Holzinger's bi-factor system (9), into principal components as by Hotelling's (10) and Kelley's (11) systems, into overlapping general and group factors as by Thurstone's system (15), into a positive general and positive and negative bipolar general factors as by Burt's system (1).

In general the system and the given matrix themselves determine, at least approximately, the number of factors, so that a hypothesis can be approximately tested as to its truth regarding the number of factors it supposes to be operative. In the typical multifactor analysis, however, fixing the number of factors still leaves free choice as to the nature of factors, over a very wide range of possible patterns. The mathematical solution leaves an infinite series of rotations possible; though, it is true, not all psychologically conceivable

factor patterns i.e. hypothetical constructs, would lie in this series. But in general, unless the prior hypothesis is very precisely defined—with quantitative characterizations far more complete than are commonly risked or attained by clinicians and others in describing their hypotheses—the search for real functional unities by using factor analysis as a confirmation of hypotheses is a farce.

Turning, therefore, to the second of the above methods, one finds, after having made a choice of analysis system, that the whole problem then resolves itself into finding principles to determine the rotation of axes. For seeking the factors in personality traits in general, in all the Protean forms in which they are likely to appear, the only entirely acceptable system seems to be Thurstone's multifactor centroid analysis. Burt's bipolar system is a special case—a special rotation or, rather, lack of rotation—from this system. None of the other systems is sufficiently flexible to permit any number of both general and group factors to emerge, in any required degree of overlap and with any pattern of sign loadings that the real functional unities may require. For example, it is hard to conceive of a general positive direction of "goodness" for all personality traits, such as would produce a wholly positive general factor. Usually (4) any one trait variable is defined in bipolar fashion, by opposites, and there is no especial reason to consider one direction rather than the other to be positive. Consequently, in any representative and random set of variables dealing with personality, those analysis systems which demand first a single positive general factor can be ruled out forthwith; and so, for other reasons, can other special systems. The problem therefore narrows itself, from this stage on, to finding principles to determine rotation of axes in a multifactor centroid analysis.

Principles for Determining Satisfactory Rotations

The following seven principles are propounded for determining the rotation of axes. The first four have been explicitly stated and employed before by various workers. Most are equally applicable to orthogonal and oblique axes solutions. It seems to the present writer that the only entirely satisfactory principles are those which *are* also applicable to oblique solutions. For it is part of the general flexibility required in any factor analytic system that it shall be able to yield oblique factors. There is no guarantee that the source traits in personality, and the social, physical, genetic and physiological influences which produce them, are entirely independent. For example, the combination of social stratification and assortative mating almost certainly produce some correlation between the basically independent genetic source traits of general mental capacity and general emotional

stability (6). For all we yet know, there may very occasionally be fairly marked departure from orthogonality, and our whole system of choosing factors should be able to adapt to this.

1. *Rotation to agree with clinical and general psychological findings.* The axes are centered on some well-known syndromes or some sets of variables, each known on other than factorial grounds to be highly involved in a psychological unity. If the syndrome became established only through observations in cross-sectional studies, this procedure amounts to nothing more than putting axes through clusters, for a syndrome is then merely a correlation cluster or *surface trait*. The alignment of source traits with surface traits then manifests the radical weaknesses discussed under the cluster principle below (No. 3). However, a syndrome may be more than a cross-sectional cluster: it may have its true functional unity witnessed by developmental and other observations. In this case there is no error, but there is also no discovery, except (a) insofar as the factor reveals the influence of the source trait in new trait elements not clinically recognized in the gross syndrome, and (b) insofar as such fixation assists in the realization of principle 5 below.

2. *Rotation to agree with factors from past factor analyses.* This procedure, which has been widely resorted to, especially in the final stages of a rotation in which simple structure no longer gives clear indications, consists in rotating until as many as possible of the factors agree with factors previously established in independent researches. The factors of earlier researches have sometimes been established as single general factors, by concentrated research in one particular field, e.g., intelligence, perseveration, surgency, using tests deliberately devised to measure what on clinical or general psychological grounds is expected to form a functional unity. Consequently these more intensive and insightful researches are used to anchor the rotations of the more dispersed multifactor research. Reyburn and Taylor, who most explicitly employed this principle (before their statement of principle 3 below) urge that knowledge of past analyses be constantly present as a guide; for "ignorance . . . even if artificially induced by a neglect of what we already know, is not compensated for by mathematical methods of factor manipulation" (13, p. 59).

On the other hand, the present writer would object, this principle may merely perpetuate psychologically fallacious concepts which guided the historically earlier researches. One must either put faith in "mechanical methods," i.e., clear general principles, or else use factor analysis only as a check on hypotheses arrived at by quite independent considerations. There is no reliable middle path of the kind suggested by this "method."

3. *Rotation to put axes through the centre of clusters.* This may be done either by picking out the outstanding correlation clusters in the original correlation matrix or by considering the clusters which exist in the projection on a single plane when the number of factors is known and plotted. The following comments apply substantially to both of these related procedures.

In general, if there are two factors operating fairly evenly in a certain matrix, the noticeable correlation clusters are likely to occur in the regions of overlap of the two factors. In these regions the shared variance (communality) is higher. Such comparatively even distribution of loadings is likely to occur when the total variance is accounted for by a considerable number of factors, as in personality variables. Clearly in such circumstances a cluster is more likely to represent a region of overlap of several factors than the region of strong influence of one factor. On the other hand, with one or two factors, the high points (clusters) of the matrix may well be the variables best defining the factors. (For example, in a matrix satisfying the two-factor theory we put the axis through the centre of the most highly intercorrelating bunch of variables.) However, since both possibilities exist there is no guarantee that a salient cluster is anything more stable than a province of overlap of two or more real, functional factors.

Reyburn and Taylor (13), criticizing the "negative" outlook in Thurstone's simple structure method, which to some extent may be said to pay attention to clusters by avoiding them, by putting the *hyperplanes* through the densest regions, develop very systematically the positive procedure of seeking clusters. In this case the clusters are not the clusters in the matrix but their projections on one plane at a time in the n-dimensional space found by the factor analysis. The investigators suggest that the first factor be rotated to pass through the centre of a cluster. This factor is then dropped from the factor matrix and the remainder are rotated to put a factor through the next most salient cluster, and so on (13).

This method has grown in popularity. But it must be objected against it—in addition to the above objections—that in practice many clusters are sheer artifacts produced by the experimenter's more or less deliberately choosing obviously psychologically related tests. Frequently, moreover, he may be unaware of the existence of the important psychological variables which, if brought into the matrix, would fill the empty spaces between his clusters, demolishing their claim to individuality and diagnostic worth as determinants. Such criticisms do not necessarily apply to the attention paid by "simple structure" to clusters, for that attention is only a secondary result of a primary

aim of over-all simplicity.

4. *Rotation for simple structure.* This principle is too well known to factor analysts, through the writings of Thurstone (**15**), to require any description here. Its essential nature is discussed by contrast and comparison under principle 7 below.

5. *The principle of orthogonal additions: rotation to agree with successively established factors.* In an n-dimensional orthogonal system, if the position of $n-1$ axes is known from previous sources of evidence, the position of the nth axis is automatically established. One can begin, therefore, with tests which, apart from specific factors, measure only known factors, or even only a single known factor. "Known" means, here, "known to correspond to a real functional entity," e.g., general mental capacity, hyperthyroidism, manic-depression. By trial and error, guided by psychological insight, one then attempts to add variables to the matrix which will introduce, apart from specifics, only one new factor. When the new factor is determined a further set of variables can be added, introducing another new factor the position of which in turn becomes fixed by the earlier factors.

In this way, starting with one factor of known position, it should be possible, theoretically, by successive additions to fix the rotation of a most complex multi-dimensional factorization. Indeed, in a relatively inexplicit and planless fashion, this principle has been employed in practical research problems, as the history of establishment of factors during the past twenty years shows. For example, Garnett's "*c*" factor (**8**), later refined into the concept of the surgent temperament (**2**), was first established as the second factor in a set of variables in which Webb's general character integration factor "*w*" was taken as the prior, confirmed, functional unity.

The defect of this method is that it depends on orthogonal axes. (It also requires more experiment with variables and more tentative factorizations than the overworked factor analyst can generally afford.) Moreover, if the first factor, used as the starting point, happened to be really the least orthogonal of any factor in the system, there would ensue a considerable systematic error in the positions of all the other factors, in addition to the errors due to each departing a little from the mean orthogonal position.

6. *The principle of expected profiles: rotation to produce loading profiles congruent with general psychological expectations.* It is possible that on general psychological grounds one could validly conclude that certain *kinds* of source traits should manifest certain general forms of factor loading pattern in certain batteries of variables. One would then rotate so that the maximum number of factors would

give loading profiles, i.e., factor patterns* of the kind required. The kinds of traits most likely to have consistent, characteristic profiles are those distinguished elsewhere as temperamental, ability, and dynamic source traits (3), and these might be subdivided further according to whether they are constitutional or environmental mold traits. To make a detailed defense of particular, definitive views as to what these profiles are is not within the scope of this paper, but one may hazard that environmental mold traits, e.g., honesty, politeness (and, especially, certain abilities, e.g., certain dexterities and skills), would tend to have all-or-nothing loadings, because they are deliberately imposed by education in a few trait elements and neglected elsewhere. That is, they would appear essentially as group factors in any general sample of trait elements or as general factors showing what we might call "plateau loadings," i.e., either very high or very low loadings. By contrast, constitutional, temperamental source traits would be expected to manifest themselves more uniformly in all trait variables, as a more smooth general factor.

According to this principle, therefore, one would rotate to get profiles of loadings having relationship to the psychological nature of the source traits (factors) as shown by the nature of the trait elements in which the factor tends to appear most heavily. For example, a factor showing bigger loadings in temperament traits than any other would be adjusted in rotation to give smooth general factor loadings. Simple structure tends to give "plateau loading" profiles, but in regard to all source traits instead of to some only, as our principle would require. The profiles expected will obviously depend also on the choice of variables. If, for example, they are all of one kind, e.g., temperament variables, so that all the factors will be of one kind, this principle will give no assistance.

The Most Fundamental Principle

More extended consideration will be given to the most basic principle, as follows:

7. *The principle of parallel proportional profiles.* This begins with the same general scientific "principle of parsimony" which forms the premise for Thurstone's simple structure, but arrives at a different formulation of the meaning of the principle in the field of factor analysis. The principle of parsimony, it seems, should not demand "Which is the simplest set of factors for reproducing this particular correlation matrix?" but rather "Which set of factors will be

* The term factor pattern is used here and later to refer to a single column in a factor matrix, i.e., to only a single slice out of what is technically called a factor pattern, e.g., by Holzinger (9).

most parsimonious at once with respect to this and other matrices considered together?" This parsimony must show itself especially when the correlations emanate from many diverse fields of psychological observation, e.g., applied, social, and physiological psychology. The criterion is then no longer that the rotation shall offer fewest factor loadings for any one matrix; but that it shall offer fewest dissimilar (and therefore fewest total) loadings in all the matrices together.*

This newer formulation depends on the consideration that if real psychological functional unities exist they are bound to appear as possible mathematical factorizations in many different kinds of situations, whereas the mathematical factors which are artifacts will stand only the test of fitting the particular matrix in which they happen to appear and may not be reproducible elsewhere.

But when one asks more precisely, "What are the derivations of the two or more correlation matrices, the simultaneous agreement of which will determine the true factors?" and "What exactly is the nature of the agreement required?", the answers have to be carefully qualified.

Regarding the derivation, it is clear that to require agreement in factors and factor loadings among correlation matrices derived from the same or similar test variables on the same or similar population samples, is an empty challenge. No new source of rotation determination is introduced, for such matrices will differ only by sampling errors and there will be an infinite series of possible parallel rotations in the two or more analyses. The special and novel required condition is that any two matrices should contain the same factors but *that in the second matrix each factor should be accentuated or reduced in influence by the experimental or situational design,* so that all its loadings are proportionately changed, thereby producing, from the beginning, an actual correlation matrix different from the first.

The changes of design or circumstance which can be introduced to bring about such orderly modification are broadly of two kinds, (a) distortion of the factorization by special selection of the popula-

* Simple structure does not preclude this condition, but it does not demand it as the primary condition. Thurstone writes, "It is fundamental criterion for a valid method of isolating primary abilities that the weights of the primary abilities for a test must remain invariant when it is removed from one test battery to another test battery" (15). But this is expected to follow from simple structure, instead of conversely. Reyburn and Taylor, on the basis of their fairly extensive experience, question the correctness of this expectation (13). By interpreting a factor analysis first in regard to a certain battery alone, and then in relation to a larger battery of which the latter formed part, they found marked inconsistency (12). "If . . . we explain the sub-battery in the simplest way, . . . we make the explanation of the larger battery, including the sub-battery, more complex, if not impossible" (13).

tion, or by altering the trait variables in some systematic fashion; (b) changing the trait measurements from measures of static, inter-individual differences to measures from other sources of differences in the same variables.

The first, which seems to the present writer a less satisfactory approach, would have, on more detailed examination, the following subvarieties: (1) differential selection of the two populations with respect to those features which are likely to constitute a functional unity, e.g., more or less age-selected, normals and psychotics, males and females, hyper- and hypothyroids, etc.; (2) change in form or method of scoring of the tests, e.g., increasing the level of difficulty, administering under speeded or unspeeded conditions, administering before and after practice periods; (3) allowing the same tests to be associated with different supplementary tests in different matrices has frequently been suggested as a test of "factor invariance," i.e., of cross-checking between two matrices, but except in special circumstances would not be a determiner of rotation in the sense required here. Some of the early work of Spearman, and more recent work by Guilford, Woodrow, Wherry and others shows clearly that method (2), however, does produce modifications of individual factor emphasis while retaining recognizably the same factors.

The second source of matrix difference requires the gathering of measurements in ways not hitherto generally envisaged in factor analytic studies of personality. The present writer has reviewed these systematically elsewhere (7). The ways of gathering measurements for correlation include, over and above the usual *Static Analysis,* (2) *Incremental Analysis,* i.e., using *changes in score* of individuals through time or experimental influence; (3) *Intraindividual Mutational Analysis,* using correlations between series constituted by many successive measurements of a set of variables in one person; (4) *Group Mutational Analysis,* of common tendencies to fluctuate, and (5) *Tied Differences Analysis,* intercorrelating differences of scores in related pairs of individuals, e.g., twins; and various modifications and extensions of these designs, providing in all an entirely ample source of matrix variations (7).

The argument for "parallel proportional profiles" as a means of determining rotations now runs as follows: (a) If one is dealing with true functional unities (unitary traits) they should show themselves alike in static, mutational, incremental, intra-individual and other analyses; (b) owing to the modifying circumstances, however, the factors will be present in different amounts, so that the loadings of a set of variables $a, b, c, d, , \cdots , n$ by factor A in the first

matrix, will appear proportionately* reduced in the loadings of the same variables by factor A_1 in the second matrix. For if a factor is one which corresponds to a true functional unity it will be increased or decreased *as a whole.*

Consequently there should be a position in the rotation of the factors from the first matrix which will give, for each factor, a profile of loadings "similar," i.e., proportional or parallel, to those obtainable from some rotation of the second matrix. With respect to most of the factors the similarity may extend to identity (apart from chance sampling errors), for the changed circumstances may often alter only one factor. If *all* are identical, no guidance is available for determining the rotation, but otherwise the problem resolves itself into asking: (a) Does a rotation exist in which the factors from each matrix will give proportionalities of loadings to those in the other, and (b) Will this matching be unique, i.e., will there be only one position in both at which such proportionality is possible? The problem is like that of rotating two or more cylinders on a combination lock to find the one position in which all will "click." In the ensuing mathematical analysis we shall deal with the above questions and ask also if there are methods, other than the formidable process of trial and error, to find those positions.

This proposed method of escaping from the alleged "psychological meaninglessness" of factor analysis, and which involves the deliberate planning of two or more parallel, coordinated correlation studies in order to determine rotation, has been named, in a purely descriptive fashion, the method of *parallel proportional profiles* (simultaneous in the sense of existing for several factors simultaneously). To indicate the historical foundations from which it builds, however, and the fact that it extends to several matrices simultaneously the principle of parsimony involved in simple structure, it might equally well be called "simultaneous simple structure."

Mathematical Treatment

For a first presentation we shall consider only two dimensions and only two test variables. There should be no fundamental difficulty in extending the argument to any number of dimensions or variables, such as occur in most practical instances. Alternatively an actual example on a larger scale could be solved by considering any two variables as representative (or any two variables at one time) and

* Closer examination may require the conclusion that the loadings of A_1 are all "functions" of the loadings of A, the functions not being one of simple linear proportion; but throughout this article, for simplicity of a first exposition, we shall assume that all the loadings of A_1 are the same linear function of the loadings of A in the same, or corresponding, variables.

perhaps by rotating in two dimensions at one time. If we proceed to attack the problems in proper logical sequence, it would seem to be necessary first to prove that sets of loadings, taken at random, on two axes from two matrices, cannot normally be brought into this special relation by rotation, i.e., that we are dealing with a special condition which cannot be simulated. Thereafter we shall attempt to prove that in the special conditions when a position *does* exist satisfying the required relations, rotation will not produce other positions equally satisfactory.

In Diagram 1 let the given, unrotated factor axes from the first matrix be x and y, and those from the second ξ and η; and let the given loadings of two tests be represented by the coordinates of A and B on the first coordinate system, and of C and D on the second coordinate system. It is required to prove that if these are random values, not satisfying any particular condition, they cannot be rotated to give the special relationship required.

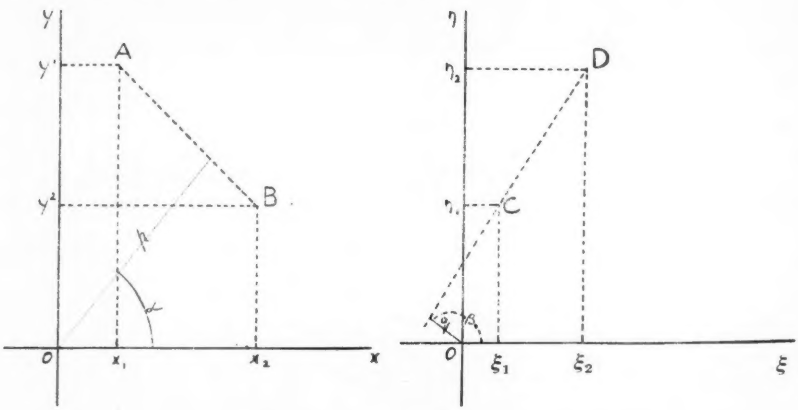


DIAGRAM 1

For simplicity we shall suppose that the special relationship existing in the special case when the factors are psychologically real is one in which only *one* factor is proportionately reduced while the other retains the same loadings in the two situations, i.e., we want to know whether a position can be found which gives the *same* loading on one axis and proportionate (parallel profile) loadings on the other.

Let us suppose that a rotation θ of the xy axes, and a rotation ϕ of the $\xi\,\eta$ axes will make the abscissas of A,C equal, those of B,D equal, and the ratio of the ordinates of A , C equal to the ratio of the ordi-

nates B,D. This rotation (with a translation to superpose the two figures on common axes) would result in the positions indicated in Diagram 2, in which the new axes are x', y'.

If the four points are not collinear in this position (and if neither line coincides with the x' axis) the lines AB, CD will be concurrent at some point on the x' axis as shown in this diagram.

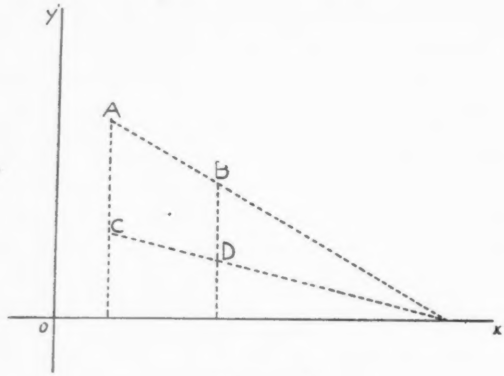


DIAGRAM 2

The conditions for rotations to be able to bring about this position are that the following three equations should hold simultaneously.

$$x_1 \cos\theta + y_1 \sin\theta = \xi_1 \cos\phi + \eta_1 \sin\phi\,, \tag{1}$$

$$x_2 \cos\theta + y_2 \sin\theta = \xi_2 \cos\phi + \eta_2 \sin\phi\,, \tag{2}$$

$$\frac{p}{\cos(\alpha-\theta)} = \frac{q}{\cos(\beta-\phi)}. \tag{3}$$

The constants p, q, α and β in the third equation are dependent upon the given variables as indicated in Diagram 1, and the equation is derived from the equations of lines AB and CD with respect to the original axes, which are as follows:

$$x\cos\alpha + y\sin\alpha - p = 0\,,$$
$$\xi\cos\beta + \eta\sin\beta - q = 0\,.$$

Since we have three equations in two unknowns these equations are in general inconsistent. The conditions of consistency can be found by eliminating θ from the first two equations, then solving for ϕ and substituting in the third. The result is complicated unless special artifices are used.

The inconsistency of the three conditions on θ, ϕ for random values of $x_1\ y_1$, $x_2\ y_2$, $\xi_1\ \eta_1$, $\xi_2\ \eta_2$ (i.e., the impossibility of simultaneously solving the equations for any two points on one set of coordinates and any two points on another) shows that in general the factor loadings of two variables in one matrix (with respect to two factors) cannot be rotated to bear the special relation here required to the factor loadings from another matrix taken at random.

The possibility of such a rotation to the required kind of congruency must therefore arise only as a result of highly specialized conditions in the data. But the special condition which we suppose to exist in the data when true functional entities are operative—namely, that in which the factor loadings (projections) are the same on one axis and proportional on another (with respect to the other matrix) is not the only one which will satisfy the above equations. It is necessary to digress, in the interests of mathematical rigor, to consider these unforseen alternatives, for examination will show that, from the point of view of psychological work, they may be definitely set aside, leaving only one solution which is both mathematically and psychologically satisfactory.

The first special situation, that in which both the lines *AB* and *CD* (in the original coordinate systems and, of course, after rotation) happen to go through the corresponding origins of the coordinate systems, as shown in the (double, superposed) Diagram 3, is one in which the three equations can be satisfied by an infinity of rotation positions.

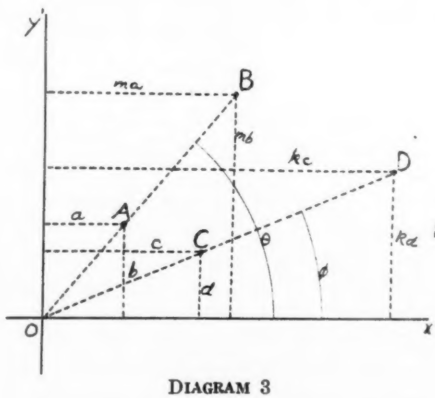


DIAGRAM 3

The proof, briefly, is as follows. Rotate, as before, one line through θ and one through ϕ. The conditions for equal x's are:

$$a \cos\theta + \quad b \sin\theta = \quad c \cos\phi + \quad d \sin\phi\,, \tag{4}$$

$$ma \cos\theta + mb \sin\theta = kc \cos\phi + kd \sin\phi\,. \tag{5}$$

A necessary condition for these being simultaneously true is that $m = k$, or $\phi = 90°$ and $\theta = 90° + \beta$. We suppose $x'y' \neq 0$. If, moreover, this necessary condition is satisfied, the only other condition is (4). This, however, can be satisfied in an infinite number of ways, as follows: Through A draw an arbitrary line RS, as in Diagram 4. Through the origin draw the axis x' perpendicular to RS.

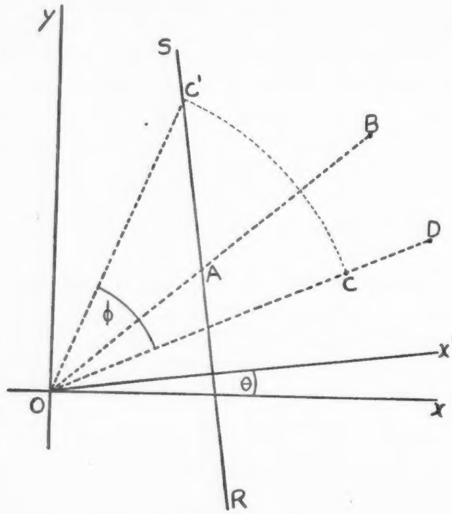


DIAGRAM 4

The angle θ is fixed by A and RS. Rotate the line OC till C falls on RS (at C'). The angle ϕ is thereby fixed (RS gives the direction of the y' axis). Since (4) is satisfied, all three conditions are now met.

Fortunately the alignment of factor loadings with the origin would only be a very rare, accidental occurrence in factor analysis and one which could be avoided by choosing, for purposes of rotation, variables which do *not* stand in this relation, so it may be eliminated from consideration.

A second special case is that in which AB, CD when placed as in Diagram 2 are symmetrical with respect to the x' axis. In such a case, an infinity of other solutions is given by $\theta + \phi = 0$; that is, if the axes are rotated the same amount in opposite senses, the segments remain symmetrical with respect to the new x-axis.

It now remains to prove that in the special case where the three equations of the main theorem above are simultaneously satisfiable there will be one and only one solution (apart from the above exceptions, namely, solutions at 180° to the unique solution, and solutions through symmetry or collinearity with origins).

Reverting to Diagram 1 we wish to show that if $x_1 = \xi_1$, $x_2 = \xi_2$, $\eta_1 = \lambda y_1$, and $\eta_2 = \lambda y_2$ the only values of θ and ϕ are $\theta = \phi = 0°$ and $\theta = \phi = 180°$ (excluding, as stated above, the case of points collinear with the origin, or the case $\lambda = \pm 1$, i.e., sheer identity of loadings, and also $\lambda = 0$, i.e., no loadings in one factor).

Substitution of the above values in (1) and (2) gives

$$x_1(\cos\theta - \cos\phi) + y_1(\sin\theta - \lambda\sin\phi) = 0\,,$$
$$x_2(\cos\theta - \cos\phi) + y_2(\sin\theta - \lambda\sin\phi) = 0\,.$$

The determinant of this system is not zero. Hence the conclusion reached is

$$\cos\theta = \cos\phi\,, \quad \sin\theta = \lambda\sin\phi\,. \tag{6}$$

From the first, $\sin\theta = \pm\sin\phi$. Substitution in the second gives

$$\sin\theta(\lambda \pm 1) = 0\,.$$

Hence $\sin\theta = 0°$ and $\theta = 0°$ or $180°$. From the first of (6) $\theta = \phi$.

But the 180° solution simply reverses the direction of the factors and does not create any new psychological solution, so that, psychologically, the solution remains absolutely unique. In short, if we can rotate to a position in the factors from one matrix where the test projections on one axis are equal to and on another proportional to the loadings from a rotation of axes from a second matrix, this position is completely defined and unique.*

The determination of this unique position is possible by solving the simultaneous equations (1), (2), (3) above, in which θ and ϕ are the (unknown) required, rotation angles.

Extension of Argument

The above theorem has dealt with two test variables and two dimensions. To be of general utility in the solution of factor analytic problems it needs to be extended as follows: (1) to apply to any number of variables; (2) to apply to any number of factors; (3) to apply to paired matrices in which not one but all factors have their contributions changed in the second matrix situation, i.e., in which the sec-

* The writer wishes to express his great indebtedness to Professor J. M. Thomas of the Duke University Mathematics Department and Miss A. Schuettler of the Wellesley College Mathematics Department for independent statements of the present solution in rigorous mathematical form.

ond matrix has all loadings proportional to, rather than equal to, those of the first; (4) possibly also to yield equations soluble for more than two matrices simultaneously, which might be computationally convenient.

If the same real psychological functional unities are at work in two matrices, then proportionality will exist equally with respect to all variables when the right rotation is obtained. Consequently if proportionality holds for two variables it will hold for all. One would therefore calculate the required rotation from further pairs of variables, after the first pair, only in order to get an average free from the chance errors particular to any one pair; i.e., there would be no systematic differences.

The extension required by (2), namely, the demonstration that the unique position can be equally well determined in data in n dimensions, seems to promise no special difficulty and will be attempted in a later article.

The extension required by (3) may not work out so satisfactorily. The requirement of proportionality rather than equality introduces a new degree of freedom. It adds a new equation without adding a new unknown. It is conceivable that with two variables and several dimensions the position in one rotation which gives loadings having, for each factor, a certain ratio to the loadings for the corresponding factor, is not uniquely determinable. With many variables, requiring a solution for any one pair which agrees with the solution for each and all of the other pairs, however, a restriction is added which may determine the position uniquely. The method of parallel proportional profiles requires for its effective use that all, or all but one, of the factors be different in their mean loading of the same variables in the two situations. Our proof has been given for the situation of differences in all but one factor, which happened also to be the situation of difference in one factor only. Experimentally, the design of having all* factors different in emphasis in the two situations is probably far easier to achieve, because it requires less control. This extension to deal with the general case in which the loadings of the variables on every factor of matrix A are functions (differing for each factor) of the loadings of corresponding variables and factors in matrix B, is therefore one which we shall investigate mathematically at the first opportunity in an ensuing contribution.

The fourth extension probably offers no essentially new prob-

* In practice it might prove difficult to ensure that all factors but one are precisely, or even substantially, unchanged in their influence in the two situations from which the matrices are calculated. It would be, perhaps, even more difficult to find a circumstantial test evidencing, before the factor analysis, that the intention to keep their influences unchanged has been realized.

lems. It requires only that, in the interests of convenience, one should find reasonably brief routine methods of solving these equations, if possible for more than two matrices simultaneously, for matrices of appreciable size and complex factor composition. For if this method indeed proves to be one way out of the impasse of indeterminacy and psychological meaninglessness, in which many psychologists seem to consider the early promise of the factor analytic attack to have bogged down, it can immediately be applied to resuscitate a considerable array of factor matrices lying in past publications.

BIBLIOGRAPHY

1. Burt, C. The factors of the mind. London: University of London Press, 1940.
2. Cattell, R. B. Temperament tests: I. Temperament. *Brit. J. Psychol.*, 1932, 23, 308-329.
3. ———. The description of personality: I. Foundations of trait measurement. *Psychol. Rev.*, 1943, 50, 559-594.
4. ———. The description of personality: II. Basic traits resolved into clusters. *J. abn. & soc. Psychol.*, 1943, 38, 476-506.
5. ———. The description of personality: III. Principles and findings in a factor analysis of the personality sphere. (Publication to be announced).
6. ———. The cultural functions of social stratification: I. Regarding the genetic bases of society. *J. soc. Psychol.*, 1944, 20.
7. ———. The operational determination of trait unities and modalities. *Brit. J. Psychol.*, (Publication date to be announced).
8. Garnett, J. C. M. General ability, cleverness and purpose. *Brit. J. Psychol.*, 1919, 9, 345-366.
9. Holzinger, K. J., and Harman, H. H. Factor analysis: a synthesis of factorial methods. Chicago: University of Chicago Press, 1935.
10. Hotelling, H. Analysis of a complex of statistical variables into principal components. *J. educ. Psychol.*, 1933, 13, 223-298.
11. Kelley, T. L. Essential traits of mental life. Harvard Studies in Education, Vol. 26. Cambridge, Mass.: Harvard University Press, 1935.
12. Reyburn, H. A., and Taylor, J. G. Some factors of intelligence. *Brit. J. Psychol.*, 1941, 31, 259-270.
13. ———. On the interpretation of common factors: a criticism and a statement. *Psychometrika*, 1943, 8, 53-64.
14. Spearman, C. Abilities of man. London: Macmillan, 1932.
15. Thurstone, L. L. The vectors of mind. Chicago: University of Chicago Press, 1935.

PSYCHOMETRIKA—VOL. 9, NO. 4
DECEMBER, 1944

INDEX FOR VOLUME 9

AUTHORS